Tampere University

Janne Lounasvaara

Foreign or Regular?

PLURAL FORMS OF THE NOUNS *ANTENNA*, *FORMULA*, *CRITERION* AND *PHENOMENON* IN BRITISH AND AMERICAN ENGLISH

Faculty of Information Technology and Communication Sciences
Master's Thesis
September 2019

TIIVISTELMÄ

LOUNASVAARA, JANNE: Foreign or Regular? Plural Forms of *Antenna*, *Formula*, *Criterion* and *Phenomenon* in British and American English
Pro gradu -tutkielma
Tampereen yliopisto
Englannin kielen ja kirjallisuuden maisteriopinnot
Syyskuu 2019

Tämä korpuspohjainen pro gradu -tutkielma tarkastelee englannin kieleen latinasta ja kreikasta lainattujen substantiivien *antenna*, *formula*, *criterion* ja *phenomenon* monikkomuotoja britti- ja amerikanenglannissa. Tutkielma selvittää miten eri monikkomuodot jakautuvat semanttisten erojen perusteella sekä kahden kielivarieteetin välillä. Semanttisten erojen lisäksi kiinnitetään huomiota kirjallisten lähteiden ja korpusaineiston yhteneväisyyteen sekä muihin korpusanalyysin aikana ilmenneisiin huomionarvoisiin havaintoihin.

Tutkimusaineiston pääosan muodostavat The Corpus of Global Web-based English (GloWbE) -korpuksesta haetut esimerkkilauseet, jotka sisältävät yhteensä 1885 yksittäin manuaalisesti analysoitua monikkomuotoa. Lisäksi tutkimusaineistona käytetään kielioppikirjoja, kielenhuolto-oppaita sekä sanakirjoja, joiden avulla muodostetaan korpusaineiston analyysissä käytettävät semanttiset tai muut tarpeelliset kategoriat.

Tutkielman teoriaosa käsittelee lainasanojen historiaa, kielimuotojen johdonmukaistumista sekä tutkimuksen kohteena olevien monikkomuotojen ilmenemistä kielioppikirjoissa, kielenhuolto-oppaissa ja sanakirjoissa. Tutkielman metodiosa tarkastelee korpuslingvistiikkaa sekä esittelee analyysissä käytettävät menetelmät. Varsinainen analyysiosa koostuu korpusaineiston analyysistä, ja tutkielman viimeisen osan muodostavat pohdinta ja johtopäätökset.

Tutkimus osoittaa, että eri substantiivien monikkomuodoilla on omat ominaispiirteensä, jotka vaihtelevat tapauksittain. *Antenna* -substantiivin osalta yleisin monikkomuoto on säännöllinen *antennas,* kun taas *formula* -sanan osalta se on vierasperäinen *formulae*. Lisäksi säännöllisten ja vierasperäisten monikkomuotojen välillä on monia yksityiskohtaisia semanttisia eroja. *Criterion* ja *phenomenon* -substantiivien vierasperäisten monikkomuotojen käyttö yksiköllisinä sanoina on huomattavan yleistä, kun taas niiden harvinaisemmat monikkomuodot, kuten säännölliset *criterions* ja *phenomenons* ovat suhteellisesti erittäin harvinaisia. Tutkielma pääosin vahvistaa kirjallisissa lähteissä esitetyt monikkomuotoja koskevat toteamukset, joskin kirjallisten lähteiden yksityiskohtaisuudessa on huomattavia eroja eivätkä ne ennusta tarkasti kaikkien monikkomuotojen ilmenemisen yleisyyttä. Lisäksi tutkimus osoittaa, että analyysissä käytetty GloWbE -korpus sisältää huomionarvoisia puutteita.

Avainsanat: englannin kieli, substantiivit, lainasanat, monikko, korpuslingvistiikka

Table of Contents

1 Introduction

The English language has several ways of expressing grammatical number. The most frequent one involves adding the suffix *-s* or *-es* to the end of the word. This is commonly referred to as the regular plural and it is the most frequent and recognized of the English plural forms. In addition, there are a handful of different types of irregular plural forms. For example, Biber et al. (1999: 286-288) list four types of "native" irregular plurals. These may include a vowel change in the middle of the word (e.g. *foot – feet*) or other additional modifications to the word (e.g. *child – children*). However, Biber et al. also list as many as six different Latin and Greek plurals that occur in English. Consequently, there are quite a few different morphological changes that English words utilize to express grammatical number. Furthermore, there are semantic and contextual differences to consider between these different forms.

This thesis examines four loanwords of a specific kind: they all have kept their original foreign plural forms in English but also occur, to varying extent, with the regular English *-s* plural. Two of the loanwords chosen for this study are of Latin (*antenna*, *formula*) and two of Greek (*criterion*, *phenomenon*) origin. Thus, each of these loanwords has at least two alternative plural forms (*antennae – antennas, formulae – formulas, criteria – criterions, phenomena – phenomenons*). The words themselves were mostly (3 out of 4) selected on the basis of my earlier bachelor's thesis topic, which in turn originated out of the realization that the complexities of alternative foreign and regular plural forms is something an average English learner does not really encounter, at least not during Finnish basic and secondary education. The number of examined lexemes was limited to four with the intention of maintaining a manageable amount of data for corpus analysis. Further contributing factors to the selection of these lexemes include their relatively common occurrence in English compared to some other words with foreign plurals (e.g. *amoeba*) and the formal similarity between the two Latin, as well as the two Greek, words.

A situation where a language user has to choose between alternative plural forms is bound to create confusion and errors and is therefore linguistically interesting. The apparent lack of research that is specifically focused on loanword plurals or foreign plural**s** and their co-existence with regular English plural forms is the key motivation behind my thesis.

This study is restricted to examining two varieties of English: British and American (henceforth BrE and AmE). The main motivation behind concentrating on these language varieties is their almost equal numerical representation in the Corpus of Global Web-based English (GloWbE), discussed in detail in Section 6.2. The aim of my thesis is to answer the following research questions:

1. What is the distribution of the plural forms of *antenna*, *formula*, *criterion* and *phenomenon* in terms of semantics and between British and American English?
2. Is the language usage data from GloWbE corpus consistent with how the plural forms of these nouns are described in grammars, dictionaries and language usage guides?
3. What other relevant observations can be made based on the corpus analysis?

To answer these questions, a considerable number of individual language usage instances (tokens) will have to be analyzed individually – a total of 1885. The GloWbE corpus is used as the primary source of language usage data on the plural forms of the nouns.

The primary literary sources of this study consist of grammars, language usage guides and dictionaries, which will provide information on the nouns that is used in establishing a meaningful categorization for the corpus data analysis. I have chosen a corpus-based approach for this study, which means that the electronic corpus is above all a method of obtaining real language usage data to be analyzed.

This study can be divided into four main parts. Firstly, Sections 2 – 5 form the theory or literature part, which discusses the topics of loanword history, regularization and the plural forms in question as they appear in the literary sources examined. Secondly, Section 6 contains the methodology part, which introduces the field of corpus linguistics and how it is used in my thesis to

perform the corpus analysis. It also addresses issues related to the scientific method. Thirdly, Section 7 forms the analysis part, which presents the corpus analysis of the 1885 tokens representing the plural forms of the four nouns studied. And finally, Sections 8 and 9 respectively provide further discussion and conclusion on the findings of the study.

My thesis does not include a hypothesis based on earlier literature and tested against the corpus data, but the aim is rather to make relevant observations of the corpus data and reflect those observations in relation to the literary sources. In this way the present study can contribute to the understanding of how accurately grammars, usage guides and dictionaries portray the reality that exists in the language usage data in GloWbE corpus.

2 Latin and Greek nouns and loanword history in English

2.1 Declensions

The four nouns chosen for the present study have their origins in the classical languages of antiquity: Latin and Greek. In their respective source languages, as well as in English, the words have similarities in terms of the morphological endings in the singular: *antenna – formula*, *criterion – phenomenon*. Latin and Greek nouns used to follow an inflectional system of three grammatical genders (masculine-feminine-neuter) and between five (Greek) and six (Latin) grammatical cases.

In the Latin case system, the first declension is sometimes called the 'a-declension' due to the nominative singular ending of its mostly feminine nouns (Jacobs 2009: 1). The nominative plural ending in the first declension is -*ae*. Thus, the original plural forms of the two Latin nouns of this study are *antennae* and *formulae*. Table 1 below illustrates the Latin first declension.

Table 1. Latin first declension

<table>
<tr><td rowspan="2"></td><td colspan="2">Aqua, aquae 'water'</td></tr>
<tr><td>Singular</td><td>Plural</td></tr>
<tr><td>Nominative</td><td>Aqua</td><td>Aquae</td></tr>
<tr><td>Vocative</td><td>Aqua</td><td>Aquae</td></tr>
<tr><td>Accusative</td><td>Aquam</td><td>Aquās</td></tr>
<tr><td>Genitive</td><td rowspan="2">Aquae</td><td>Aquārum</td></tr>
<tr><td>Dative</td><td rowspan="2">Aquīs</td></tr>
<tr><td>Ablative</td><td>Aquā</td></tr>
</table>

In Greek, *criterion* (κριτήριον) and *phenomenon* (φαινόμενον) are part of the second declension, also called the 'o-declension' according to the "stems to which the case endings are attached" (Smyth 1956: 47). For nouns that are neuter in their grammatical gender, the Greek second declension has the plural ending -α. Accordingly, the original nominative plural forms of the two nouns are *criteria* (κριτήρια) and *phenomena* (φαινόμενα).

It should also be mentioned that the English noun system itself has changed considerably during centuries. As pointed out by Baker (2012: 50), Old English had several major and minor declensions (and grammatical cases), whereas in terms of plural forms, modern English has only one major declension, the *-s* plural, and a few minor ones (e.g. the *-en* plural in *oxen*). According to Fischer et al. (2001: 72), the "whole-sale simplification" of the original Old English system had made the regular *-s* plural dominant by the 15th century.

2.2 Loanword history

This section presents a brief and general overview on loanword history in English in relation to Latin and Greek borrowing. The more detailed examples of the first attested use of the actual loanwords examined in this study will be discussed in Section 5.4.

The current situation of Latin and Greek loanwords in English is summarized by Durkin (2014: 6) as follows:

> …more formal language in modern English and/or more academic topics of discussion generally involve using a higher proportion of borrowed words than more casual everyday conversation. These are chiefly words borrowed from French and/or Latin, or words formed ultimately from elements that come from Latin or Greek.

This fairly obvious statement reflects the historical development in which the classical world extended itself across centuries in the form of language of literacy and institutions such as the Catholic Church and academia.

Given the fact that English as a separate language of the Germanic language family did not yet exist during classical antiquity, the earliest Latin borrowings still present in English were probably taken over during proto-Germanic times (ibid. 72). The estimated total of Latin-derived vocabulary, compounds and derivatives included, in Old English is around 4-5% (ibid. 100). However, it was the Norman Conquest, beginning in 1066, that resulted in much more significant changes in the nature and structure of English vocabulary. Borrowing from French, a descendant of

Vulgar Latin itself, reached its zenith in the first half of the 14th century, although in many cases it cannot be established with certainty whether a word is from French or Latin. A combined origin is likely for many (ibid. 236).

Durkin adds that the height of Latin borrowing into English, in terms of absolute numbers of new words, occurred in the 16th and 17th centuries and increasingly so that the Latin words were restricted to formal or scientific registers (ibid. 299). According to van Gelderen (2014: 179), English borrowed many words from Latin and Greek during the Renaissance because of a lack of suitable terms required at that time. She quotes Görlach (1991: 136), who asserts that the period from 1530 to 1660 witnessed the fastest expansion of English vocabulary in the history of the language. Such an expansion was presumably aided by the printing press, a somewhat new innovation during the Renaissance. Thus, the expansion of Latin and Greek loanwords in English was motivated by the need to express ideas and concepts that spread during the early modern period. The fact that existing English words, or new English-based coinages, were not chosen to carry out this task presumably reflects the firmly established role of Latin and Greek as the languages of science in the past, but also the prestige still carried by them.

The entrance of Greek loanwords into English requires transliteration from one alphabet to another. Therefore, "most of the Greek words have entered into English through Latin, or have, at any rate, been Latinized in spelling and endings before being used in English" (Jespersen 1912: 114).

To summarize, Latin and Greek loanwords have entered into English mainly via Latin, some via French. The defining characteristic of these borrowings is that the loanwords are very much related to certain types of registers, especially formal and scientific ones, as opposed to loanwords from other source languages. This can be seen as a consequence of the historical developments in European science and culture, which are closely intertwined with the Greco-Roman culture and its rediscovery in early modern times

3 Problematic co-existence of foreign and regular plurals

Many grammars of English use 'foreign plural' either as a sub-category of 'irregular plural' or 'plural', such as Declerck (1992), Huddleston et al. (2002), Leech & Svartvik (2002), and Quirk et al. (1985), or use another specific (sub)classification, such as 'Latin and Greeks plurals' (e.g. Biber et al. 1999).

In a similar manner, I use the term 'foreign' to refer to the original Latin or Greek plural and 'regular' to the English *-s* plural. 'Irregular' may denote either an irregular Old English-derived plural (e.g. *children*) or an irregular foreign plural, which of course is irregular only from the English point of view and regular in its source language.

According to Huddleston et al. (2002: 1590), a persistent problem with foreign plurals is that there is no way of inferring a correct form from the base of the word. For example, final *-a* is characteristic of one class of Latin nouns (the first declension mentioned in Section 2.1), but also such words as *algebra* (from Arabic) and *phobia* (from Greek). Quirk et al. (1985: 305) add that whereas it is helpful to know about pluralization in relevant source languages, such knowledge is still unreliable because some loan words do not conform to the original plural patterns (e.g. *areas*, *villas*) while others do (e.g. *larvae*).

In other words, an English user cannot always be familiar with various inflectional paradigms affecting different - sometimes superficially similar - loanwords, nor the intricacies that have come to determine the use of different plural forms. For instance, the originally Latin plural form *data* has become disassociated from its original singular *datum* and is often treated as both singular and plural (Biber et al. 1999: 287). This unpredictability is a key problem when it comes to a language user's choice between alternative plural forms.

According to Burchfield (1996: 442), there is a shift towards regular plurals with some loanwords (e.g. *referendums* instead of the original *referenda*), aided by the fading knowledge of Latin. On the other hand, there is a further comment that "the choice of plural form sometimes

depends on the subject area" (ibid.). This means that alternative plural forms of the same word can be associated with different contexts and have separate meanings. It is not unreasonable to think that such differentiation contributes to the survival of foreign forms that otherwise carry the burden of Latin or Greek inflections in English. This view is supported by Crystal (2009: 249), who, in a discussion on the alternative adjectival endings *-ic* and *-ical*, present the "desirable tendencies" of 'differentiation' and 'clearing away the unnecessary':

> When two forms coexist & there are not two senses for them to be assigned to, it is clear gain that one should be got rid of
>
> (ibid. 250)

Garner's (2003: 615) view is that:

> Many imported words become thoroughly naturalized; if so, they take an English plural. But if a word of Latin or Greek origin is relatively rare in English – or if the foreign plural became established in English long ago – then it typically takes its foreign plural.

This seems to be in contradiction with McMahon's (1994: 73) claim that frequency is what actually protects irregular forms from regularization. Garner does not discuss why a Latin or Greek loanword would become established in the first place. In the previous section some possible explanations were brought forward, i.e. fulfilling a terminological void and bringing along the prestige required in a particular register.

Peters (2004: 314) remarks that the oldest loans from Latin, such as *cheese* and *oil*, have completely assimilated, whereas the later arrivals tend to have the foreign form at least alongside the regular. She also notes that "Latin loanwords which are strongly associated with an academic field usually have Latin plurals as well" (ibid. 2). Thus, a firm association with a register or a clear semantic specialization would presumably account for the survival of foreign plural forms when most of the Old English system has been decimated by the Modern English regular plural. The corpus data analysis section of this study will explore the issue of the distribution between form and meaning, to a certain extent.

The plural forms examined in my thesis have been subject to different prescriptive guidelines by grammarians and lexicographers. Writing over 90 years ago, Ball (1928: 296-314) summarizes the then dictionary treatment of the alternative plural forms as follows:

- *antennae – antennas* -> only the foreign form is given
- *formulae – formulas* -> both foreign and regular forms are given, regular is preferred
- *criteria – criterions* -> both foreign and regular forms are given, foreign is preferred
- *phenomena – phenomenons* -> only the foreign form is given

As I will demonstrate later in this study, these guidelines do not seem to quite fit with modern usage data and guidelines, for various reasons. Ball does not provide any justification for the preferences between these plural forms as he merely describes the status quo of his time and place. Garner's (2003: 615) general advice is to choose the regular form when in doubt, so as to avoid hypercorrection or overregularization. His message seems to be that hypercorrection can cause more harm, perhaps unintelligibility, when applied to irregular forms.

On the basis of these views, several different factors affect how loanwords preserve or lose their original plural forms. No general rule that fits all instances can be given, and there is no agreement on which factors are more defining than others. Further recommendations or preferences for "correct" plural forms expressed in usage guides and dictionaries will follow in Section 5. The following section approaches the topic of regular *-s* plurals from the perspective of regularization.

4 Processes of regularization

4.1 Regularization

The English noun system has developed into one that, in terms of frequency, strongly favors the regular plural over a handful of minor irregular plurals, such as the foreign plurals borrowed from Latin and Greek. The present study is particularly interested in co-existing alternative plural forms. The fact that some nouns have adopted the regular *-s* plural alongside an earlier foreign form is part of a phenomenon known in linguistics as regularization. This section discusses the processes of regularization, particularly those of analogy, on the basis McMahon's (1994) work on language change.

Regularization is a common process in languages. As the term suggests, it means replacing irregular forms (e.g. morphological elements like plural endings) by regular ones. Regularization has been documented extensively in children's language acquisition, formation of creole languages and sign languages, and in historical trends of language change (Ferdinand et al. 2019: 53). Earlier studies on regularization have dealt with these areas of language but there seems to be a lack of research when it comes to the specific problematics of foreign versus regular plural forms. According to Zapf and Ettlinger (quoted in Warfelt 2012: 178), the regular *-s* plural:

> …is one of the earliest learned grammatical morphemes in the English language, appearing in children's productions as early as 18 months of age (de Villiers and de Villiers, 1973; Zapf and Smith, 2007), but not showing complete mastery until as late as seven years of age (Berko, 1958).

Since Latin and Greek nouns have been borrowed into English usually to be used in formal registers, as was discussed in Section 2.2, it is only natural that they are not a central research topic in the context of child language acquisition. However, the early emergence of the regular plural in child language development is interesting. It may reflect some of the reasons why that particular form came to dominate the English plural system: something about it, perhaps its "easiness", favors

its adoption. Morphologically the regular plural is certainly easier than the complex system of different inflectional paradigms which it has replaced to great extent.

With regard to phonology, Zapf and Ettlinger (Warfelt 2012: 178) elaborate by dividing the regular plural form itself into two codas: simple and complex. The former signifies the *-s* morpheme after a vowel (vowel + consonant) and the latter in a consonant cluster, such as in *dogs*. Research indicates that simple coda forms emerge earlier than complex ones. In this sense, morphological simplicity would favor the emergence of regular plural forms such as *antennas*, *formulas*, *criterions* and *phenomenons*, out of which phonological simplicity would further encourage the first two. This is of course a crude simplification and does not take into account many other forces at play, for instance semantics.

Regularization is closely related to the processes of analogy. McMahon (1994: 70) presents analogy as a "housekeeping device" that creates regularity where irregularity has been produced, often due to sound change. According to her, the task of analogy is to keep three types of structures in line: sound structure, grammatical structure and semantic structure. In relation to the expansion of the regular English plural, two subtypes of analogy are worth discussing here.

4.2 Analogical extension

Analogical extension is the generalization of an already existing morpheme or relation into new forms or situations (McMahon 1994: 71). The Modern English application of the regular *-s* plural is a clear example of analogical extension. McMahon remarks that the complex Old English (OE) system had no way of signaling merely grammatical number but noun inflections also carried information about gender and case, and so did adjectives, pronouns and the definite article. There were different inflectional paradigms, i.e. combinations of suffixes and modification to the noun stem, none of which was dominant over the others. McMahon (ibid.) illustrates the situation with the inflectional paradigm of the OE noun *stān* 'stone', shown in the table below:

Table 2. Old English declension of *stān* 'stone'

	Singular	**Plural**
Nominative	stān	stānas
Accusative	stān	stānas
Genitive	stānes	stāna
Dative	stāne	stānum

Already in OE, a regularization of an earlier more complex inflectional paradigm had taken place. As Table 2 shows, the earlier distinctions between nominative and accusative forms within singular and plural have disappeared, which was not the case in older Germanic languages like Gothic (ibid. 72). With *stān*, only the final /s/ proved stable enough an inflectional ending to be reinterpreted as a marker of plural and genitive and to be analogically extended to many other nouns which previously did not include an /s/ in their paradigms. Analogical extension is frequently observed in child language as children overregularize forms such as *foot* into **foots* instead of irregular *feet* (ibid.).

Even though there are highly successful analogical extensions like the regular *-s* plural, analogy is rarely exceptionless. There are still irregular plurals in Standard English although the processes of regularization have been at work since the time of OE. McMahon (ibid. 73) offers frequency as an explanation as to why some irregular forms have avoided regularization tendencies, such as analogical extension, for so long. If an irregular form occurs frequently, it is also susceptible to being corrected, for example when a child is learning a language and produces incorrect forms. On the other hand, there is evidence that irregular forms are acquired before regular ones, at least when it comes to verbs, as pointed out by Marshall and van der Lely (2012: 126).

Again, no one explanation accounts for all the peculiarities. The plural form *oxen* has arguably not been very frequently used in recent decades but it has resisted the analogical extension

of the regular plural nevertheless. Regardless of these kinds of exceptions, there seems to be a general connection between analogy and frequency (McMahon 1994: 73).

4.3 Analogical levelling

Another systematic type of analogy is called analogical levelling. McMahon (1994: 73) distinguishes analogical extension and analogical levelling so that the former involves patterns whereas the latter has to do with paradigms, i.e. sets of inflectional forms with the same stem morpheme. Analogy in general is connected to sound change and analogical levelling exhibits this connection by levelling, i.e. removing, the opaqueness that a sound change may have caused within a paradigm of a verb or a noun.

McMahon exemplifies analogical levelling with the words *sword* and *swore* (ibid. 74). Whereas a sound change caused the formerly pronounced /w/ to disappear between /s/ and a back vowel in *sword*, analogical levelling restored it in *swore*, which makes the paradigm of the verb *swear* more coherent. In other words, analogical levelling interferes with sound change but does not reverse it completely.

4.4 Predicting analogy

For the purposes of this study, it is important to try to generalize some of the reasons why regular plural forms have been adopted alongside foreign ones in the first place. According to McMahon (1994: 77), among the main sets of generalizations made about analogy are Kuryłowicz's six laws. Kuryłowicz's fifth law states that:

> …if the speakers of a language have a choice between keeping a contrast of rather marginal significance, and abandoning it in favour of reinstating a more basic distinction, then they will abandon the marginal contrast and reestablish the basic one.
>
> (ibid. 78)

The quote above relates to regular English plurals in the sense that the /s/ marker was chosen to stay in use by English speakers when morphological markings of case were falling into disuse. As the /s/ marker, adopted from the declension illustrated in Table 2 earlier, became analogically extended to

be the marker of grammatical number in other noun paradigms too, the importance of marking the basic distinction between the singular and plural strengthened its position. In line with this, McMahon (ibid. 80) presents a summary of Kuryłowicz's laws and Mańczak's (both Polish linguists) tendencies on how analogy is predicted to operate. Among the predictions is the elimination of multiple expression of the same information. The overwhelming adoption of the plural marker /s/ and the disappearance of most of the other forms expressing plural would support this prediction, although counterexamples can often be found.

4.5 Iconicity and Humboldt's Universal

There are two further notions that should be mentioned in the context of regularization and analogy. McMahon (1994: 85) describes the principle of iconicity so that it "seems to favour related surface elements which are similar in form as well as meaning, and which more generally binds language to the non-linguistic world". In other words, the reduction of the numerous plural markers of OE to almost exclusively *-s* would entail that *-s* has taken on the meaning 'plural' and a shift from an arbitrary sign towards an icon would have occurred. However, the form and meaning of the marker are not purely isomorphic (i.e. one-to-one) because of its role as the marker of genitive as well.

There are contentions that human language would be conceptually ideal if one form always corresponded to one meaning but this conceptual ideal is in conflict with phonetic ideals and therefore interrupted by sound change (ibid. 90). McMahon (ibid. 91) presents an 'innate principle of linguistic change' called Humboldt's Universal by Vennemann (1978: 259). It claims that grammatical markers should be unique and constant. This is consistently the case with children's regularization of irregular forms, e.g. noun plurals, during language acquisition (ibid.).

If we accept that there is an innate tendency in language change towards iconicity and reduction of redundancy, and this tendency is carried out by processes of regularization, such as analogy, then it should also manifest itself in the plural forms of the nouns in this study. A prediction would then be that there is semantic differentiation between foreign and regular plural

forms: one meaning is connected to one form. I will return to this question in the following sections when examining the plural forms as they appear in literature and corpus data.

In this section, I have discussed the processes of regularization, which account for the emergence of regular plural forms alongside original foreign forms with many loanwords. McMahon's (1994) views draw from previous studies on language change and therefore should not be considered an all-encompassing explanation for all things related to alternative plural forms. Nevertheless, I will refer to the terminology presented in this section for practical purposes later in this study.

The frequent occurrence of the English regular plural is the result of analogical extension that began during OE when language change began eliminating the grammatical expression of case and the former multiple noun paradigms. Regularization may be motivated by a general requirement of iconicity that is somehow innate in language. The next section moves on to examine the presentation of the different plural forms of *antenna*, *formula*, *criterion* and *phenomenon* in earlier literature.

5 Plural forms of *antenna*, *formula*, *criterion* and *phenomenon* in earlier literature

5.1 Introduction

The primary literary sources I have selected for this study to provide information on the alternative plural forms consist of five grammars, eight language usage guides and ten dictionaries. As pointed out earlier, there seems to be a lack of research concerning the occurrence of these plural forms, so it is necessary to resort to sources of this type. I will use these literary sources to help formulate a categorization for the corpus data analysis discussed further in the methodology Section 6. Some of the sources have already been quoted in earlier sections in the context of their general views on foreign and regular plurals. This section discusses the particular plural forms examined in my thesis. These literary sources are listed in the References section under the separate heading 'primary literary sources.'

5.2 Grammars

5.2.1 *Antennae – Antennas*

Biber et al. (1999: 287) give the following account:

> Both regular and irregular plurals are found with *antenna* and *formula*, but the irregular forms are predominant in both cases (though only regular forms were instanced in the conversation texts of the LSWE Corpus)

It is notable that Biber et al. base their grammatical description on the 40 million-word *Longman Spoken and Written English Corpus* (LSWE) (ibid. 4). Declerck (1992: 63) also recognizes both plural forms but observes a distinction: "antennae (*of an animal or insect*)/antennas (*Am. E*) (*aerials*)". Huddleston et al. (2002: 1591) allow either. According to Leech & Svartvik (2002: 359) "*antennas* is found in general uses and in electronics […] but *antennae* in biology". For Quirk et al. (1985: 311), it is simply a noun with both plurals.

5.2.2 *Formulae – Formulas*

Biber et al., Declerck, Huddleston et al. and Quirk et al. (ibid.) treat these plural forms exactly as those of *antenna* by merely acknowledging the existence of both alternatives. Leech & Svartvik (ibid.) provide more information: *formulas* is found in "general use" and *formulae* often in mathematics.

5.2.3 *Criteria – Criterions*

Biber et al. (ibid. 288) only mention the foreign plural but add that occasionally, rarely, it is used as a singular. Declerck (ibid. 64) and Leech & Svartvik (ibid.) only accept *criteria* as the plural. Huddleston et al. (ibid. 1593) regard the foreign plural as correct but also note "very rare examples" of the regular form, and a more common but not widely acceptable use of *criteria* as a singular. According to Quirk et al. (ibid. 312), the foreign plural is common but there is an irregular, widely condemned use of *criteria* as a singular and *criterias* as a plural.

5.2.4 *Phenomena – Phenomenons*

Biber et al.'s (ibid.) description is similar to the one in 5.2.3: the foreign form is the correct one but it is occasionally used as a singular. Just as above, Declerck and Leech & Svartvik (ibid.) only accept *phenomena* as a plural. For Huddleston et al. (ibid.) the foreign form is correct, and it has an occasional, not widely acceptable, use as a singular. Quirk et al. (ibid.) note that whereas the foreign plural is the norm, *phenomena* sometimes occurs informally as a singular.

5.2.5 Summary

In summary, two out of five grammars report a difference, semantic or register-related, between *antennae* and *antennas*. One does the same with *formulae* and *formulas*. All grammars favor *criteria* and *phenomena* as the correct plurals and mostly do not even mention their regular forms. Three out of five grammars acknowledge a rare use of *criteria* and *phenomena* as a singular, and one observes the possible occurrence of *criterias*.

5.3 Usage guides

5.3.1 Introduction

The selection of usage guides consists of five British (Burchfield 1996, Crystal 2009, Howard 1993, Swan 2005 and Peters 2004) and three American guides (two Garners 2003 & 2016 and Davidson 2001). Peters (ibid. vii) has drawn much of her data from two corpora: the British National Corpus (BNC) and the Cambridge International Corpus of American English (CCAE). The publication of these guides spans from 1993 to 2016, but it should also be mentioned that H.W. Fowler's original *A Dictionary of Modern English Usage*, published in 1926, forms the core of Burchfield (1996) and Crystal (2009). The usage guides are in no way uniform but vary greatly in their level of precision, from a non-existent entry for a word, or a simple list of correct plural forms, to extensive commentary on different aspects affecting the usage of a word.

5.3.2 *Antennae – Antennas*

Davidson (2001: 39) assigns the regular form to "sending and receiving radio waves" and the foreign to insect organs or metaphoric use relating to human alertness. Burchfield (1996: 50) agrees with the semantic distinction between the two forms but also suggests the foreign form to be more common in BrE (ibid. 36). According to Garner (2016: 54), the current ratio in favor of *antennae* when referring to insects, as opposed to the regular plural, is 17:1 and that of *antennas* when referring to devices is 4:1. Howard (1993: 25) also endorses this basic semantic distinction. Peters (2004: 40) claims that there is a more than 90% preference for *antennae* in biological and figurative use.

In Section 3, I presented Ball's summary on dictionary usage from 1928, which did not accept the regular *antennas* at all. A logical explanation would be that radio antennas were still rare, and the innovation of assigning the regular plural specifically to them had not yet happened. It

seems that the later emerged *antennas* has found its own niche and according to the principle of iconicity (Section 4.5) the two different plural forms refer to different things.

Apart from Swan (2005), who does not list *antenna* at all, the language usage guides' basic rule is: *antennae* for biological referents and figurative use, *antennas* for technical devices.

5.3.3 *Formulae – Formulas*

Howard (1993: 174) states that the foreign plural is more likely to be used in scientific contexts. He also admits that the "now accepted" regular plural is usual elsewhere. Garner's (2006: 407) view is very similar in his statement that the regular plural "predominates in all but scientific writing". Crystal's (2009: 190) claim is that *formulae* and *formulas* are equally common. However, the guide also includes *formula* among the words whose plural forms vary according to context, the regular being preferred in popular writing, the foreign in "scientific treatises" (ibid. 316). The earlier original statement in Fowler's work was that in AmE, both plural forms were "reported to be equally common in all senses" (Burchfield 1996: 310). This differs from Peters' (2004: 217) view that AmE would be almost wholly behind the regular plural, apart from contexts of scientific and scholarly writing. She also refers to a ratio of 3:1 from the British National Corpus displaying evidence for BrE preferring the foreign plural *formulae*. Swan (2005: 517) lists both plurals in his examples but does not discuss any distinctions in meaning or between varieties.

5.3.4 *Criteria – Criterions*

Crystal (2009: 400) counts the noun among Greek-derived ones that "often or always" have the foreign plural. He adds (ibid. 754) that in speech or "unmoderated written language" it is increasingly common to use *criteria* as a collective singular noun. In Burchfield (1996: 191), the foreign form is the correct plural and also often erroneously used as a singular. Garner (2006: 233) treats *criteria* as the only correct plural but admits that sometimes the non-standard *criterions* occurs and that "[i]nfrequently, though not infrequently enough, one even sees **criterias*." He also

claims that especially from around the mid-20th century there have been attempts to "make criteria a singular" (ibid.).

Howard (1993: 107) advises using the foreign form, as does Swan (2005: 524). Peters describes the foreign form as standard and also provides frequency information with regard to the word's standard singular and plural forms. According to her:

> **Criterion** is in fact the less common of the two, outnumbered by **criteria** by more than 1:3 in the BNC and almost 1:4 in CCAE. Thus **criteria** is far more familiar for many, a fact which helps to explain its increasing use as a collective singular noun.
>
> (Peters 2004: 133)

She goes on to make a relevant comparison to the nouns *data* and *media*, as does Crystal (2009: 754), which "are also now construed in collective and singular senses" (ibid. 134).

5.3.5 *Phenomena – Phenomenons*

Davidson (2001: 351) recommends the foreign form and admits that the regular exists but deems it unnecessary and unappealing. Garner (2016: 689) too favors the foreign plural and describes it as erroneous to use *phenomena* as singular or the regular plural *phenomenons*, with an exception:

> But in the popular sense "a talented person who is achieving remarkable success and popularity", *phenomenon* makes the plural *phenomenons*.

For Howard (1993: 311), *phenomena* is the correct plural, sometimes erroneously used as singular. Swan (2005: 517) lists *phenomena* as the one and only plural form.

Peters (2004: 420) is more elaborate and explains that the word's plural form has been causing trouble in English from the very start. According to her, the confusion persists partly because "phrases like *natural phenomena* and *psychic phenomena* often seem to be collective concepts, rather than countable plurals" and that assimilation of the singular and plural is more advanced in AmE. She mentions the existence of the regular *phenomenons* in the sense of "outstanding person" but also cites an example of *phenomena* used in that sense. Peters' conclusion

is that the foreign form is securely dominant over the regular for plural uses while at the same time extending its use as a singular.

5.3.6 Summary

Compared to the cautious indications given by the grammars, the usage guides almost unanimously express a clear and strong semantic division between *antennae* (e.g. insect organs) and *antennas* (technical devices). The usage guides also agree that *formulae* and *formulas* have a semantic or contextual division between roughly 'scientific/formal' and 'other/general'. There is also indication of BrE preference for the foreign and AmE preference for the regular form.

As for *criteria* and *criterions*, the message is that the foreign form is the preferred standard, the regular form is a rare exception and the use of the foreign form as a singular is relatively common. The analogous plural *criterias* is mentioned in two usage guides. The description of *phenomena* and *phenomenons* is somewhat similar: the foreign form is the endorsed standard but it is used as a singular to the extent that a development towards widely accepted singular status may be underway. Contrary to the plural forms of *criterion*, a semantic distinction between *phenomena* and *phenomenons* is expressed in some usage guides.

5.4 Dictionaries

5.4.1 Introduction

The dictionaries selected for this study include four general-purpose dictionaries (two British, two American), three collegiate dictionaries (American) and three learner's dictionaries (British). Thus, there are five British and five American dictionaries. These dictionaries will be referred to by abbreviations given below in square brackets.

The British general-purpose dictionaries are: *The Oxford Dictionary of English* (3rd edition) [ODE] and *The Oxford English Dictionary - OED online* [OED]. The American general-purpose dictionaries are: *The American Heritage Dictionary of the English Language* (5th edition) [AHD]

and *Webster's Third New International Dictionary* [W3]. The collegiate dictionaries are: *The American Heritage College Dictionary* (4th edition) [AHC], *Merriam-Webster's Collegiate Dictionary* (11th edition) [MER] and *Webster's New World College Dictionary* (5th edition) [WCD]. And the learner's dictionaries are: *Cambridge Advanced Learner's Dictionary* (4th edition) [CAM], *Collins COBUILD Advanced Learner's Dictionary* (8th edition) [COL] and *Oxford Advanced Learner's Dictionary* (9th edition) [OAL]. The *OED online* entries for *antenna* and *phenomenon* are updated 3rd edition entries. The dictionaries were published between 2003 and 2017, except for W3, which was published in 1961 with an Addenda Section last updated in 2002.

Due to the large number of short quotations, italicization in the sections below will be used to highlight a) the plural forms, both within and without the quoted dictionary passages, and b) dictionary definitions or senses when quoted in short phrases. Bolded numbering represents the numbering of the different senses of the words given in the dictionaries.

5.4.2 *Antennae – Antennas*

OED provides the first attested use of the foreign plural in English as follows:

> 1646 Sir T. Browne Pseudodoxia Epidemica iii. xviii. 153 Insects that have *antennæ*, or long hornes to feele out their way, as Butter-flies and Locusts.

There is a very consistent pattern of description for this pair of alternative plural forms in the dictionaries. Firstly, the terminology differs slightly but the general agreement is that *antenna* refers to a *sensory appendage* (OED), *feeler* (WCD), *insect or crustacean part* (COL) or *sensory organ* (MER). AHC and AHD use the categorization *zoology* followed by a detailed description:

> One of the paired, flexible, segmented sensory appendages on the head of an insect, myriapod, or crustacean functioning primarily as an organ of touch.

Secondly, eight out of ten dictionaries agree that in this sense the plural of the noun is *antennae,* with the exception that WCD and W3 also allow *antennas* in this case and OED provides one example of such use. Furthermore, seven out of the ten dictionaries also recognize a figurative or metaphorical use of *antennae* (or *antennas* [CAM]) to signify *the faculty of instinctively detecting*

and interpreting subtle signs (ODE), as in: "The minister was praised for his acute political *antennae*" (OAL).

The third point of general agreement among the dictionaries is associating the plural form *antennas* with technical devices. According to OED, *antennas* is the plural especially in the technical sense. OAL is the only dictionary not making any distinction between the use of *antennae* and *antennas* when the referent is a technical device. Along the lines of the usage guides earlier, there is a great deal of unanimity as regards the use of these two plural forms in the dictionaries studied.

5.4.3 *Formulae – Formulas*

All the dictionaries recognize the co-existence of *formulae* and *formulas* but only a couple make a distinction regarding their usage, or give any rules for it. According to OAL, *formulae* is used *especially in scientific language*. Similarly, ODE assigns the foreign plural to mathematical and chemical expressions and the regular to other uses.

The OED entry includes 4 different senses. There are a total of fourteen example sentences of the headword in the plural. These split evenly between seven *formulae* and seven *formulas* but no guidelines are given for any semantic or contextual differentiation. OED's first attested use of the word is an example in the singular from 1583[1], representing sense **1a**:

> A set form of words in which something is defined, stated, or declared, or which is prescribed by authority or custom to be used on some ceremonial occasion.

There is significant variation in the dictionary definitions of *formula*. CAM lists only two separate senses for the word: **1** *a method/rule* and **2** *baby's milk*. The other extreme is OAL with its seven definitions, which are:

[1] 1583 A. Nowell et al. True Rep. Disput. E. Campion sig. Ee2v Camp... The Formula of the second couenant, is Christ. Charke. You vnderstande not..what Formula is.

> **1** (mathematics) a series of letters, numbers or symbols that represent a rule or law **2** (chemistry) letters and symbols that show the show the parts of a chemical compound **3** a particular method of doing or achieving sth **4** a list of the things that sth is made from, giving the amount of each substance to use **5** (also formula milk) (especially NAmE) a type of liquid food for babies, given instead of breast milk **6** a class of racing car, based on engine size, etc **7** a fixed form of words used in a particular situation

The word is obviously used in a variety of ways and it is not always clear how to draw lines between the different senses. Consider, for example, senses **1b** and **4** of MER: where does *a conventionalized statement* end and *a customary or set form or method* begin?

For most of the senses in all dictionaries, both plural forms are given as equal alternatives, or at least there are no stated restrictions, an indication that the plural forms are not significantly divided in their meaning, which is allegedly the case with *antenna*. Only two dictionaries describe a separate usage for the two plural forms.

5.4.4 *Criteria – Criterions*

In OED, the first quoted examples date to the early 17th century when the word occurs written in Greek alphabet within otherwise English sentences, the first Latin alphabet instance being from 1661. The regular plural occurs in one of the example sentences, dated 1788.

Six out of the ten dictionaries (AHC, AHD, MER, OED, WCD and W3) recognize both foreign and regular plural forms. In addition, AHC, AHD, ODE and WCD mention a wide and often objected use of *criteria* as a singular. According to AHC, this use is *not yet acceptable*, whereas with the analogous plurals *agenda* and *data* it is.

Criterion differs from the three other studied words in that there is no sign of any semantic differentiation between the alternative plural forms. The dictionaries define the word as having one or two senses, which are essentially the description found in AHC: “A standard, rule, or test on which a judgment or decision can be based.” Unlike some of the usage guides, none of the dictionaries refer to the occasional occurrence of the form *criterias*.

5.4.5 *Phenomena – Phenomenons*

Like the other words studied, *phenomenon* is first recorded in English in the late 16th – early 17th century (1583 in OED). The dictionaries vary between one and four senses in their definitions. Perhaps surprisingly, OED lists three different plural forms in its entry: *phenomena*, *phenomenons* and *phenomenas* and also provides examples of each occurring from early on. The regular form *phenomenons* is mentioned in seven out of ten dictionaries. The regular form has the following associated definition in AHD:

> **2.** *pl.* **-nons a.** An unusual, significant, or unaccountable fact or occurrence; a marvel. **b.** A remarkable or outstanding person; a paragon. See Synonyms at **wonder**.

The foreign *phenomena* is always the first given plural form in the dictionary entries and it is generally defined in a very broad sense as having one to three senses. For instance, COL lists only one sense: "A *phenomenon* is something that is observed to happen or exist". Some dictionaries are more precise in their definitions and, for example, separate the senses into *phenomena* of physics and those in Kant's philosophy (senses **3** and **4**, AHD).

A notable feature about the OED entry is that it lists a total of seventeen different spellings for the three plural forms. These include spellings with varying first vowels, such as *phainomena*, or even apostrophes (e.g. *phoenomena's*). This is significant for the corpus analysis and will be revisited in the methodology Section 6.3. MER, ODE and OED mention that a singular use of *phenomena* exists but is not standard or formal.

5.4.6 Summary

It is now evident that the four nouns have their specific differences and characteristics. *Antenna* is the most obviously semantically divided in its plural forms. The semantic divergence of *phenomena* and *phenomenons* is also mentioned. The plural forms of *formula* are allegedly more interchangeable, with some indication of difference in their preferred use.

Criterion is semantically non-polysemous. *Formula* is the most polysemous of the words, with the largest number of senses. An interesting observation is the widely reported singular use of both *-a* ending Greek plurals *criteria* and *phenomena*. The dictionaries do not refer to differences in use between AmE and BrE, unlike some usage guides. All these words have been first recorded in English within a rather short time span: 1583 – 1646.

6 Methodology

6.1 Corpus linguistics

The present study is constructed around the central role of language data retrieved from an electronic corpus, and thus falls within the domain of corpus linguistics. As Tognini-Bonelli (2001: 1) points out, the debate whether corpus linguistics should be defined as a theory or methodology has existed from early on. Given the fact that corpus linguistics has authentic data as the starting point certainly makes it an empirical approach (ibid. 2). I have chosen a corpus-based research approach for my thesis. What this means is explained by McEnery and Hardie (2012: 6) in the following words:

> Corpus-based studies typically use corpus data in order to explore a theory or hypothesis, typically one established in the current literature, in order to validate it, refute it or refine it. The definition of corpus linguistics as a *method* underpins this approach to the use of corpus data in linguistics.

Another main approach in corpus linguistic studies is called 'corpus-driven'. Taken to its extreme, this approach rejects previous hypotheses of language and claims that they can only be drawn from the corpus data itself (ibid.). In other words, the two approaches can simplistically be reduced to corpus-as-theory (corpus-driven) and corpus-as-method (corpus-based) (ibid. 153).

Taking into account the fact that I use previous literature to make assumptions about the corpus data (by using already existing categorizations), places this study on the corpus-based side of these two approaches. However, McEnery and Hardie's own view (ibid. 6) is to reject this binary distinction on the basis that the corpus in itself has no theoretical status and in this sense all corpus

linguistics is more or less corpus-based. In this study, the corpus is first and foremost a method of obtaining large amounts of language data to answer the research questions. In the broad sense, corpus-based linguistics is "any approach to language that uses corpus data and methods" (ibid. 241).

A defining feature of corpus linguistic studies is the combined application of quantitative and qualitative analysis. While large data sets can easily be accessed and quantified in a split second due to advanced electronic corpora, it is often not enough for an insightful analysis. For example, in the previous section I illustrated some of the differences that manifest themselves in dictionaries and usage guides arising from different interpretations or points of view on semantics made by lexicographers. Language is simply too diverse and vague to be pinned down and explained exhaustively. The convergence of quantitative and qualitative methods is therefore almost unavoidable. As summarized by McCarthy (2015: xi):

> Yet it goes without saying, plausible interpretation and qualitative judgments informed by the statistical data are the ultimate test of the worth of any applied corpus linguistic enterprise

In this study, the quantified corpus data is qualitatively measured against the data from primary literary sources. My approach is not so much testing a hypothesis, but rather making observations in a descriptive manner and drawing conclusions from the data. The next section introduces the primary source of language data in my thesis, the electronic corpus.

6.2 The Corpus of Global Web-based English (GloWbE)

According to the GloWbE corpus website[2], the corpus is composed of 1.9 billion words on 1.8 million web pages from 340,000 websites in 20 different English-speaking countries. The corpus is what McEnery and Hardie (2012: 6) refer to as a 'balanced or sample corpus', in other words:

[2] https://www.english-corpora.org/glowbe/

"A careful sample corpus reflecting the language as it exists at a given point in time, is constructed according to a specific sampling frame." The web pages were collected in December 2012 using a process which is explained in detail on the website[3]. This means that the language data represents language found on websites up until then, but there is no indication of how far back in time the oldest included websites reach.

I decided to use the GloWbE corpus for three reasons. Firstly, it allows the comparison of different language varieties conveniently with one and the same user interface. Secondly, it contains an almost equal representation of language data from the two largest varieties, which also presumably represent similar registers. Thirdly, as mentioned above, it is a sample corpus which means that the language data is steadily the same and the same searches and results can presumably be retrieved over and over again.

The corpus contains 387.6 million words in the BrE and 386.8 million in the AmE subsection (the corpus uses the abbreviations GB and US). These were gathered from 381,841 web pages on 64,351 web sites for BrE and 275,156 US web pages on 82,260 sites for AmE. The corpus uses a classification that divides these web sites into 'general' and 'blogs' but this feature is not particularly useful or relevant for this study, nor does it affect the corpus data analysis.

A GloWbE corpus search provides source information about the search item (e.g. a word) which includes the country of the web page, the page's title and a source link to the web page. There is also a passage of text, 'expanded context', presenting the search item in its context, which may prove useful in case the link to the original web page is dysfunctional or obsolete. In this way the GloWbE does not provide the entire text as such but a possibility to view a glimpse of it and follow a link to its source, which is a convenient way to avoid copyright issues.

When viewed in the 'context' tab, the search results are listed in ascending numerical order with the web page country, source address and some of the 'expanded context' passage visible as

[3] https://corpus.byu.edu/glowbe/help/textsm.asp

well. This search result view is significant when determining which search result tokens belong to which sources, a topic dealt with in the next section, and important in checking that the analysis of search results is not skewed by multiple instances from a single source.

6.3 Methods used in the corpus data analysis

There are several matters that need to be taken into account in the corpus data analysis. Most importantly: 1) what to search, 2) how to limit the data, 3) how to classify the data/what information to look for in it, and 4) how to ensure accountability, falsifiability and replicability. This section explains the methods I have used to address these issues.

6.3.1 Search words

In the first research question, I presented the aim to find out the numerical and semantic distribution of the plural forms of the four different nouns in two language varieties. In Section 5, it was demonstrated that for some of the nouns there are more than just two different forms to consider. Furthermore, in the case of *phenomenon*, there are several different attested spellings (cf. OED) to take into account. All possible misspellings of the nouns are so numerous that including them would be impractical and outside this study. Hypothetically, there is also a small possibility that a singular form of some of the nouns would occur in the data used as a plural but because there are no such recorded instances in the literary sources, that possibility was not further investigated. Only such word forms as discussed in Section 5 are included.

Based on the literary sources and the principles stated above, Table 3 below displays the search words used in this study. Only the search words that returned at least one token as a result are included in this study and therefore listed in the table.

Table 3. Token distribution in GloWbE

Search word	Number of tokens in BrE	Number of tokens in AmE
Antennae	236	240
Antennas	279	363
Formulae	599	265
Formulas	697	1313
Criteria	11542	9255
Criterions	5	11
Criterias	16	8
Phenomena	3624	4370
Phenomenons	39	34
Phenomenas	3	1
Phaenomena	10	18
Phainomena	1	1
Phoenomena	1	0

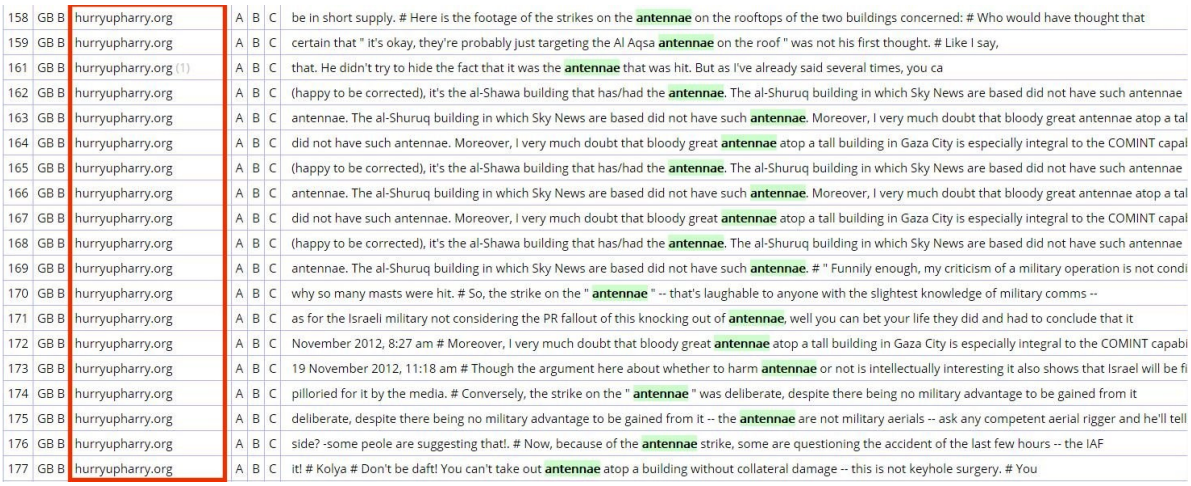
These numbers represent the total number of tokens for the search words in the two subsections of the corpus, not the number of analyzed tokens, which must be limited for practical reasons. As can be seen, there is great variation in frequency between different word forms, somewhat less so between the same forms in the two varieties. Of the 17 different spellings listed in OED for the plural forms of *phenomenon*, six returned search results.

6.3.2 Limiting the data

The number of analyzed tokens must be kept within reasonable boundaries, as there is no point in analyzing semantically unambiguous *criteria* 20,797 times. It would also be beyond the scope of this study to include several hundreds of tokens per search word considering the amount of time needed to analyze them individually. I decided to limit the maximum number of tokens to be analyzed to 150 per search word per language variety. Any figure below that is included entirely. When applied to the numbers in Table 3, the total number of analyzed tokens adds up to 1948.

However, there are more issues that need to be addressed. The most crucial one has to do with the distortion of analyzable data. Consider the picture below:

Picture 1. Screen capture of GloWbE search results for *antennae* (BrE) in context tab



158	GB B	hurryupharry.org	A	B	C	be in short supply. # Here is the footage of the strikes on the **antennae** on the rooftops of the two buildings concerned: # Who would have thought that
159	GB B	hurryupharry.org	A	B	C	certain that " it's okay, they're probably just targeting the Al Aqsa **antennae** on the roof " was not his first thought. # Like I say,
161	GB B	hurryupharry.org (1)	A	B	C	that. He didn't try to hide the fact that it was the **antennae** that was hit. But as I've already said several times, you ca
162	GB B	hurryupharry.org	A	B	C	(happy to be corrected), it's the al-Shawa building that has/had the **antennae**. The al-Shuruq building in which Sky News are based did not have such antennae
163	GB B	hurryupharry.org	A	B	C	antennae. The al-Shuruq building in which Sky News are based did not have such **antennae**. Moreover, I very much doubt that bloody great antennae atop a tal
164	GB B	hurryupharry.org	A	B	C	did not have such antennae. Moreover, I very much doubt that bloody great **antennae** atop a tall building in Gaza City is especially integral to the COMINT capal
165	GB B	hurryupharry.org	A	B	C	(happy to be corrected), it's the al-Shawa building that has/had the **antennae**. The al-Shuruq building in which Sky News are based did not have such antennae
166	GB B	hurryupharry.org	A	B	C	antennae. The al-Shuruq building in which Sky News are based did not have such **antennae**. Moreover, I very much doubt that bloody great antennae atop a tal
167	GB B	hurryupharry.org	A	B	C	did not have such antennae. Moreover, I very much doubt that bloody great **antennae** atop a tall building in Gaza City is especially integral to the COMINT capal
168	GB B	hurryupharry.org	A	B	C	(happy to be corrected), it's the al-Shawa building that has/had the **antennae**. The al-Shuruq building in which Sky News are based did not have such antennae
169	GB B	hurryupharry.org	A	B	C	antennae. The al-Shuruq building in which Sky News are based did not have such **antennae**. # " Funnily enough, my criticism of a military operation is not condi
170	GB B	hurryupharry.org	A	B	C	why so many masts were hit. # So, the strike on the " **antennae** " -- that's laughable to anyone with the slightest knowledge of military comms --
171	GB B	hurryupharry.org	A	B	C	as for the Israeli military not considering the PR fallout of this knocking out of **antennae**, well you can bet your life they did and had to conclude that it
172	GB B	hurryupharry.org	A	B	C	November 2012, 8:27 am # Moreover, I very much doubt that bloody great **antennae** atop a tall building in Gaza City is especially integral to the COMINT capabi
173	GB B	hurryupharry.org	A	B	C	19 November 2012, 11:18 am # Though the argument here about whether to harm **antennae** or not is intellectually interesting it also shows that Israel will be fi
174	GB B	hurryupharry.org	A	B	C	pilloried for it by the media. # Conversely, the strike on the " **antennae** " was deliberate, despite there being no military advantage to be gained from it
175	GB B	hurryupharry.org	A	B	C	deliberate, despite there being no military advantage to be gained from it -- the **antennae** are not military aerials -- ask any competent aerial rigger and he'll tell
176	GB B	hurryupharry.org	A	B	C	side? -some peole are suggesting that!. # Now, because of the **antennae** strike, some are questioning the accident of the last few hours -- the IAF
177	GB B	hurryupharry.org	A	B	C	it! # Kolya # Don't be daft! You can't take out **antennae** atop a building without collateral damage -- this is not keyhole surgery. # You

The area I have surrounded with a red rectangle is the part in the context tab which displays the web page sources of the tokens, which are highlighted with green. The picture illustrates that the consecutive tokens 158-177 of *antennae* in BrE all come from the same web page, and even the same text. The total number of *antennae* in BrE being 236, of which 150 will be included in my analysis, it would not be methodologically sound to allow 19 consecutive tokens from the same source to distort the data.

To avoid such distortions, I will include in my analysis only the first token that appears in a group that visibly has the same source (i.e. the same text on the same web page) and discard the rest. This means that the analyzable tokens will not be those from numbers 1 to 150 but there will be gaps, and to reach 150 tokens the last analyzable item can be, for example, token number 229. The same policy of taking into account one token from one source is naturally followed with the search words that resulted in fewer than 150 tokens as well.

The fact that the GloWbE corpus interface itself does not automatically exclude multiple tokens from the same source is one of its shortcomings, which are revisited later on at the end of Section 8. Consequently, the total distortion-corrected number of analyzable tokens in this study is 1885.

6.3.3 Classifying the data

In addition to establishing what items to search and how many, it is necessary to establish what to do with the search results. This means that a search result token must be interpreted and categorized. As stated before, I will use a semantic classification, when applicable, based on the primary literary sources. However, not all of the nouns need to be semantically classified. The literary sources show no semantic variation between the different plural forms of *criterion*. Instead, they report a relatively common use of *criteria* as a singular form. Therefore, with *criteria* the relevant information to look for among the search results is whether it occurs in the plural or singular.

According to the literary sources, *phenomena* is similar to *criteria* with regard to its use as a singular word. Thus, the search result tokens for this word too must be screened for the information on grammatical number. Defining grammatical number is not always straightforward because there may not be a verb form or determiner present to give a clue. This is illustrated by an example phrase from Peters (2004: 420): “a clearer view of the *phenomena* they are investigating”.

Besides the issue of grammatical number, *phenomena* and *phenomenons* are described as having some degree of semantic differentiation, so the word involves at least two types of information that must be considered when going through the individual search result tokens.

In the case of *antenna*, the literary sources indicate that its plural forms are fairly strongly divided between two (or three) senses. Classification is made easier when the semantic divergence is low and clear lines between senses are expected to be observed in the search results.

Formula, on the other hand, has between two to seven definitions in the dictionaries, which means that the meaning and context of analyzable tokens requires special attention and the semantic classification depends on the level of detail chosen.

As for classifying the context where the search result tokens occur, I have decided to leave it outside this study. Context information might well be relevant with Latin and Greek nouns, as

indicated by the discussion on loanword history, and some of the literary sources. However, such an effort would be beyond the scope of this study because it would entail establishing an unknown number of different categories for vast numbers of websites where the tokens are found, and it would have to be done manually because the GloWbE itself only uses the 'general' and 'blogs' classification for the websites. Besides, if, say, *formulae* is observed in the sense of 'mathematical rules', it would be reasonable to presume a connection between the word and certain types of contexts rather than others anyway (e.g. formal/education). I will occasionally comment on individual tokens in relation to the contexts they occur in, especially when discussing the search words with low numbers of search results, but otherwise context is not part of the classifications I will use.

In summary, the two types of information I will focus on when performing the analysis of individual tokens of the corpus data relate to semantics with *antenna*, *formula* and *phenomenon* and grammatical number with *criterion* and *phenomenon*. Additionally, the third research question (see Section 1) demands that I will bring up any other observations that, by subjective estimation, are relevant to this study.

Finding out the meaning of a word largely depends on finding out the word's referent. As it can be expected that this cannot always be done with certainty, it is necessary to reserve a classification for unclear cases as well. The following classification for the corpus data analysis was established on the basis of combining the information from the primary literary sources with that gained from preliminary corpus searches:

Table 4. Classification used in corpus data analysis

Plural forms of ***antenna***	A. Zoology
	B. Device
	C. Figurative use
	D. Proper noun
	E. Multiple/overlapping
	F. Unclear
Plural forms of ***formula***	A. Scientific use (e.g. mathematical, chemical)
	B. Method of doing or achieving something
	C. Fixed set of words often used ceremonially
	D. Ingestible or applicable substance (e.g. mother's milk substitute)
	E. Motor racing
	F. Multiple/overlapping
	G. Unclear
	H. Proper noun
Plural forms of ***criterion***	A. Plural
	B. Singular
	C. Unclear
	D. Proper noun
Plural forms of ***phenomenon***	A. Something observable 1. Plural 2. Singular 3. Unclear
	B. Something exceptional 1. Plural 2. Singular 3. Unclear
	C. Proper noun
	D. Multiple/overlapping
	E. Unclear

The category 'proper noun' needed to be added to include a few tokens of the sort. The category 'multiple/overlapping' is reserved for tokens whose referent is identified but cannot be placed in one category unambiguously. 'Unclear', on the other hand, means that the referent of the word is unidentified, for example due to uninterpretable context and dysfunctional web link to the source text in GloWbE. Above all, the corpus data analysis is characterized by the necessity of going through individual search result tokens manually one by one. It is the only way to obtain the information required in this study. It also implies that the judgements made during the analysis may be open for revision in some cases.

6.3.4 Accountability, falsifiability and replicability

In their discussion on the scientific method and corpus linguistics, McEnery and Hardie (2012: 15) bring up three important notions: accountability, falsifiability and replicability. The first means that data which is favorable to the hypothesis must not be purposefully selected. In this study, it is easily avoided since there is no hypothesis per se, and the selection of analyzable tokens, as explained in Section 6.3.2, is done by the GloWbE corpus apart from the anti-distortion measure of manually discarding multiple tokens from the same source. As the authors cited later point out:

> Short of using the corpus in its totality, total accountability can in principle be preserved by using an unbiased (e.g. randomized) subsample of the examples in the corpus.
>
> (ibid.)

As regards falsifiability, again, there is no hypothesis or claim to be falsified in this study, as it aims to observe and describe corpus data and compare it to the information in literary sources. However, the qualitative analysis of the corpus data, i.e. classifying the search result tokens according to Table 4, is a subjective endeavor done by the analyst and therefore open to disagreement.

Replicability is closely related to falsifiability. The choice of the GloWbE corpus as the main source of language data serves replicability in the sense that it is a sample corpus (see Section 6.2), which means that the data remains as it is and any corpus search, when replicated, produces the same list of tokens in the same order. I have aimed to ensure the falsifiability and replicability of this study by the following procedures:

I. The search result tokens are listed in ascending numerical order on the GloWbE corpus context tab. Thus, whenever a token is used as an example in this study, it will be accompanied by the search result number and language variety information.
II. All analyzed tokens are listed in Appendices A, B, C and D at the end of this study. They present the information on token number, classification (Table 4) and language variety.

In this way, every token can be identified in connection to the number it occurs with[4] and the semantic or other classifications I have assigned to it. Furthermore, the tokens I have left outside the analysis can be inferred by inspecting whether their numbers occur on the lists given in the appendices. The next section focuses on the corpus analysis itself by presenting the numerical distributions, providing example sentences and discussing the analysis in general.

[4] This turned out not to be entirely accurate. See Section 8 for discussion on the deficiencies of the GloWbE corpus.

7. Corpus data analysis

7.1 Plural forms of *antenna*

The distribution tables in the following sections present the actualized distribution of tokens, which means that if a particular search word produced no tokens in a category (see Table 4), that category is excluded from the distribution table (with one exception later on). The individual tokens are italicized in the example sentences given, whether that is the case in the original source text or not.

7.1.1 *Antennae* in BrE

None of the tokens among the 150 analyzed had such an unclear referent as to fall into the category 'unclear', which is therefore excluded from the table below. Three tokens referred to proper nouns (two to the same: Antennae Galaxies[5]). One token has a known referent but could not be placed into any of the categories due to it overlapping multiple.

Table 5. Classification and token distribution of *antennae* in BrE

Classification	Number of tokens out of 150	Percentage
A. Zoology	51	34%
B. Device	50	33.33%
C. Figurative	45	30%
D. Proper noun	3	2%
E. Multiple/overlapping	1	0.67%

The one token in category E (token 184) is a convenient demonstration of how complex it can be to semantically classify words that occur in actual language data:

> Android fan Marc Young from Ontario, Canada has made this brilliant Android robot. It has moving arms, *antennae* and head, but most importantly it looks really really cool...

The referent of the token here is a part of a robot made to resemble the logo of Google Android operating system. So the *antennae* are not really a device, nor are they figurative in the sense of the

[5] The galaxy collision resembles an insect's antennae, which is how the pair got the name. The "antennae" are formed by two long tails of stars, dust and gas expelled from the galaxies as a result of their interaction. http://www.constellation-guide.com/antennae-galaxies/

"political antennae" of an opportunist politician. Furthermore, the green *antennae* of the Android logo might as well be those of an insect[6]. Tokens that explicitly referred to functional technical devices were placed in category B.

Otherwise the distribution is fairly even between A, B and C. It should be clarified that category C includes tokens that have the sense of "ability of interpreting subtle signs" (see Section 5.4.2). For example, token 120:

> Pupils' *antennae* will be sharper if they attend solo, but many find it useful to have another set of eyes

Thus, instances where the token's direct referent were concrete insect *antennae* or where insect *antennae* were mentioned indirectly were placed in category A, and only such figurative uses as the example above in C. For instance, an imagined phrase "he made his hair stand up like insect *antennae*" would place the token in A. Likewise, if the referent related to a fictional character with insect-like *antennae*, the token fell into category A.

It is notable that the figurative use is almost as common as A and B and the foreign plural form occurs in B perhaps with unexpected frequency, if compared to the statements found in the literary sources.

7.1.2 *Antennae* in AmE

AmE resembles BrE very closely when it comes to the frequencies between categories A and B. However, there is a prominent difference in the frequency of figurative use (C) between the two varieties: it is clearly more frequent in BrE. As with BrE, the foreign plural can refer to technical devices completely acceptably and perhaps more than expected.

[6] https://developer.android.com/distribute/marketing-tools/brand-guidelines.html

Table 6. Classification and token distribution of *antennae* in AmE

Classification	Number of tokens out of 150	Percentage
A. Zoology	60	40%
B. Device	59 (3 used as singular, 1 misspelling)	39.33%
C. Figurative	22	14.67%
D. Proper noun	2	1.33%
E. Multiple/overlapping	4 (1 used as singular)	2.67%
F. Unclear	3	2%

However, there is an unexpected discovery that deserves attention. Four tokens occurred being used as a singular and one token apparently as a misspelled *antennas*:

First off, there are not "HDTV *antennae*'s". (token 107)

The singular use of the plural forms of *antenna* was not considered relevant when formulating the classification. Nevertheless, these tokens were easily noticeable due to incongruent verb agreement or the use of an indefinite article, as with token 67: "[…] get excellent picture with a $100 HD *antennae*." The web link to the original source is dysfunctional so it is not clear whether this actually is an instance of singular use or an error in the reproduction of the original text by the GloWbE corpus in the 'expanded context' view.

The two tokens used as proper nouns had the same referent as in BrE: The Antennae Galaxies. There were also three tokens the referent of which could not be determined. The figurative use includes phrases such as "conspiracy theory *antennae*" or "faith *antennae*" (tokens 51 and 185). Category E tokens involved an overlap of A and B, possibly C. Otherwise, the two most frequent categories included fairly typical references to the insect world, on one hand and TV, internet or mobile phone equipment, on the other.

7.1.3 *Antennas* in BrE

The distribution in Table 7 below illustrates that the regular plural is almost exclusively reserved for technical devices. While the foreign plural did not by any means rule out category A, the regular plural almost does.

Table 7. Classification and token distribution of *antennas* in BrE

Classification	Number of tokens out of 150	Percentage
A. Zoology	3	2%
B. Device	144	96%
C. Figurative	2	1.33%
D. Proper noun	1	0.67%

The referent of the only token in category D (token 146) is a word in a music album title. Figurative use seems to be very rare, which could mean that the metaphorical use is closely associated with the antennae found in the animal kingdom. One of the few category A tokens (token 228) refers to a cake resembling a caterpillar:

> Create a face on your final sponge and secure it to the front, you can use the candles as *antennas* if you like

In summary, the distribution suggests that the semantic specialization of the two plural forms presented in the literary sources only concerns *antennas*.

7.1.4 *Antennas* in AmE

Compared to the previous distribution, the one below is quite similar. Figurative and zoological uses are slightly, but only slightly more frequent while category B dominates the distribution.

Table 8. Classification and token distribution of *antennas* in AmE

Classification	Number of tokens out of 150	Percentage
A. Zoology	7	4.67%
B. Device	135	90%
C. Figurative	6	4%
D. Proper noun	1	0.66%
E. Multiple/overlapping	1	0.66%

The only hint at differences between BrE and AmE so far is the more frequent figurative use of *antennae* in BrE. As for *antennas*, the corpus data does not indicate any remarkable differences. An example of category A can be found, for instance, in a passage of literary fiction:

> The windows were open and on the counter were flies, black balls with sparkling translucent wings pointing askew, little *antennas*, poor little things.

The one instance in category E (token 36) refers to the appearance of fictional children's characters (Teletubbies) and therefore overlaps at least B and C. The figurative instances involve phrases such as "weak social *antennas*" (token 245).

When all four tables in section 7.1 are put together, the two most frequent categories B and A account for approximately 85% (509 tokens) of the 600 (4x150) analyzed tokens, with 388 tokens in the former and 121 in the latter category. A regular plural referring to technical devices is without a doubt the most frequent individual occurrence of the plural forms of *antenna* in the corpus data, representing 46.5% (279/600) of the tokens in all semantic categories and both varieties.

Antennae is overwhelmingly preferred for zoological and figurative uses but it also makes up more than 1/3 of tokens referring to devices. The semantic differentiation in this sense is not as strict as suggested by the literary sources.

7.2 Plural forms of *formula*

7.2.1 *Formulae* in BrE

Some of the literary sources, mainly usage guides, advised the reader that the foreign plural is closely associated with scientific use. The corpus data supports this view to a large extent.

Table 9. Classification and token distribution of *formulae* in BrE

Classification	Number of tokens out of 150	Percentage
A. Scientific	113	75.33%
B. Method	14	9.33%
C. Fixed set of words	10	6.67%
D. Ingestible substance	2	1.33%
E. Motor racing	10	6.67%
G. Unclear	1	0.67%

Perhaps against expectations, there were no category F (multiple/overlapping) tokens among the 150 items analyzed. Amusingly enough, the only category G token (token 199) is unclear because it refers to the plural form itself and therefore does not fit into the rest of the classification:

> …formula, which may be pluralized to formulas but also *formulae*

Tokens in A refer most often to mathematical but also chemical and computer programming *formulae*. Category C tokens often related to religion, for example token 40[7]:

> In fact neither the name Muhammad itself nor any Muhammadan *formulae* (that he is the prophet of God) appears in any inscription dated before the year 691 A.D.

Tokens in B included recipes, formulae for life, "emotional and behavioural *formulae*" (token 63) or references to legal formalities. The occurrence of ten tokens in category E is somewhat unexpected, since the use of the regular plural might seem more appropriate with racing cars. For instance:

> Money plays just as big a role in junior *formulae* as in F1 (token 3)

Only two tokens witnessed the foreign plural referring to infant *formula*.

[7] The same token in the same text passage reoccurred later as token number 146, which was discarded from analysis upon noticing. See Section 8 for discussion on the deficiencies of the GloWbE corpus.

7.2.2 *Formulae* in AmE

Table 3 in Section 6.3.1 showed that the total number of *formulae* tokens in AmE is less than half of that in BrE, which indicates a significant American preference for the regular form with this lexeme. As for the semantic distribution within *formulae*, it is not drastically different from BrE, as Table 10 below illustrates.

Table 10. Classification and token distribution of *formulae* in AmE

Classification	**Number of tokens out of 150**	**Percentage**
A. Scientific	103 (2 used as singular)	68.66%
B. Method	22	14.67%
C. Fixed set of words	20	13.33%
D. Ingestible substance	3	2%
E. Motor racing	1	0.67%
F. Multiple/overlapping	1	0.67%

A scientific referent lies behind more than two thirds of the tokens. It occurs in various contexts, also very informal ones, such as:

> Comments # Looks like a math *formulae*: 1+1-1 = 1, although the population has increased and in spite of it nothing has changed and therefore it's not worth mentioning. (token 205)

The above token can no longer be found by following the original web link, but the 'expanded context' view in GloWbE reveals that it used to be found in the comment section of a web page. This erroneous use as a singular was one of the two instances that occurred.

Category B is slightly more frequent than in BrE, and so is C. It is possible that American web sites have more religious content using the word, as in the following:

> Ever since the fifth century, we have used different *formulae* to confess our common faith in the One Lord Jesus Christ, perfect God and perfect Man. (token 180)

There was also one token (202) belonging to category E, despite the fact that formula motor racing is not particularly prominent in the United States. In the only instance of category F (token 95), it could not be established whether the referent was more of B or C.

7.2.3 *Formulas* in BrE

The distribution in Table 11 shows that the regular plural is, clearly, but not overwhelmingly less frequent with scientific referents than the foreign plural.

Table 11. Classification and token distribution of *formulas* in BrE

Classification	Number of tokens out of 150	Percentage
A. Scientific	73	48.66%
B. Method	44	29.33%
C. Fixed set of words	13	8.67%
D. Ingestible substance	12	8%
E. Motor racing	6	4%
G. Unclear	1	0.67%
H. Proper noun	1	0.67%

Category B is more common with regular than foreign plural and includes phrases such as: "I do not accept any absolute *formulas* for living" (token 194). Somewhat unexpectedly, there are fewer tokens in category E than with the foreign plural in BrE. Category D may contain other referents besides infant formula, for example token 138:

> Avena Sativa is often found in conjunction with Tribulus as they appear to work via similar mechanisms and can be found in many bodybuilding *formulas*.

One token occurred in an unintelligible 'expanded context' with a dysfunctional web link to its source and one belonged to the name of a medical company.

7.2.4 *Formulas* in AmE

In terms of the total number of tokens in GloWbE, *formulas* is twice as frequent in AmE as in BrE (Table 3, Section 6.3.1). The semantic distribution of the regular plural between the two varieties is not remarkably different. AmE has a little higher preference for it in scientific use, as the table below illustrates.

Table 12. Classification and token distribution of *formulas* in AmE

Classification	Number of tokens out of 150	Percentage
A. Scientific	88	58.67%
B. Method	41	27.33%
C. Fixed set of words	10	6.67%
D. Ingestible substance	11	7.33%

The regular plural may be more acceptable in AmE as type A than in BrE, since it can occur in very formal contexts as well. Token 207 is found in a quote from *Stanford Encyclopedia of Philosophy*:

> An interpretation of a formal system *U* in a formal system *V* is given by a translation ′ of *formulas* of *U* to *formulas* of *V* that preserves provability: If $U \vdash A$ then $V \vdash A'$

Occasionally *formulas* in category B might as well be replaced with the word *methods*. This interchangeability is apparent in the sentence of token 162: "There are no fixed *formulas* or methods for finding the value of items of clothing."

In category B, the regular plural is more than twice as frequent as the foreign plural when both varieties are considered. The regular plural is also preferred in category D, although the total number of tokens is low. The combined numbers of tokens in both varieties in category A are 216 for *formulae* and 161 for *formulas*, which are not as far apart as might be expected based on the literary sources.

7.3 Plural forms of *criterion*

The classification presented in Table 4 (Section 6.3.3) is only applicable to the foreign plural *criteria* because, explicitly and visibly ending in *-s*, the other two plural forms are not confused with singular use. Therefore, *criterions* and *criterias* are discussed without a quantified classification in this section.

7.3.1 *Criteria* in BrE

Compared to the previous two nouns, the plural forms of *criterion* are straightforward to analyze because there is no semantic dispersion into several senses. The primary interest lies in

investigating the distribution between the plural and singular uses, and any unclear or peculiar tokens. The distribution for *criteria* in BrE is the following:

Table 13. Classification and token distribution of *criteria* in BrE

Classification	Number of tokens out of 150	Percentage
A. Plural	139	92.66%
B. Singular	9	6%
C. Unclear	1	0.67%
D. Proper noun	1	0.67%

Whether the 6% of tokens used as a singular form is as 'widespread' as some of the dictionaries reported (see Section 5.4.4) is debatable, but certainly the usage exists in the corpus data as well. Category A tokens often had determiners, verb agreement or other clues to enable the classification. Furthermore, a knowledge of the real world is sometimes necessary in qualitative analysis. For instance, the phrase "money lending *criteria*" requires no grammatical clues because a situation with only one *criterion* is simply absurd. One notable feature of category A tokens was their very frequent (not quantified) occurrence in different versions of the lexicalized phrase "meet the *criteria*".

At least two of the category B tokens were found in online newspaper articles inside quotes from interviewed people. For example, token 125:

> A spokesman said it was "standard procedure whilst police investigations were ongoing" and added: "If a person has been arrested on suspicion of sexual offences one *criteria* stipulates they can't be left alone with children."

Some of the singular uses apparently occurred in various web site comment sections. These two observations indicate informal and/or spoken language. One token referred to the name of a recording studio, and one fell into category C due to structural ambiguities in the context.

7.3.2 *Criteria* in AmE

Compared to BrE, category B is slightly more frequent in AmE. The preferred use is predominantly in the plural and, as before, many occurrences of the phrase "meet the *criteria*" were encountered.

Table 14. Classification and token distribution of *criteria* in AmE

Classification	Number of tokens out of 150	Percentage
A. Plural	135	90%
B. Singular	14	9.33%
C. Unclear	1	0.67%

What is interesting about the use as a singular form in AmE, is that the usage can be found in the speech of high profile individuals in more or less formal contexts. For example, token 20 occurs in a Time Magazine interview of former United States president Barack Obama:

> "So what I'm trying to do is to take the best ideas from either party, with one *criteria*, one filter, and that is: Is this helping to grow the middle class, build the middle class and create ladders of opportunity for people?"

A similar instance occurs in an NBC News interview of a U.S. senator (token 287). Otherwise, category B tokens occur in various contexts from online newspaper articles to comment sections of a blog website.

7.3.3 *Criterions* in BrE and AmE

Table 3 in Section 6.3.1 displayed a distribution of 5 tokens in BrE and 11 in AmE for the regular plural. This is radically fewer than the combined 20797 tokens of *criteria*.

If the distributions in Tables 13 and 14 are used as a guideline to conclude that the singular use of *criteria* is somewhere between six and nine percent as a whole, it would mean that even the singular use alone is approximately a hundred times more frequent than the regular plural form. *Criterions* is obviously a very marginal phenomenon. Nevertheless, it has been recorded from at least the 18th century on, as pointed out in Section 5.4.4, and among the tokens in the GloWbE corpus, there is one (BrE, token 1) dating back to that period.

Otherwise it seems that the regular plural occurs in informal contexts, usually in web site comment sections but once also in a scientific publication[8] (AmE, token 6): "The maximum iterations were set to be ten with zero convergence *criterions*."

Due to the very small number of tokens it is not certain whether the frequency difference between BrE and AmE reflects the situation more generally, or if it simply occurs coincidentally in the corpus data.

7.3.4 *Criterias* in BrE and AmE

After purging the tokens that have the same web source, the distribution of *criterias* is 14 in BrE and 8 in AmE. Both the low number of tokens and the contexts they are found in resemble those of *criterions*. This analogous plural occurs in comment sections or, for example, in a transcript of an interview (BrE, token 3): "The *criterias* have to be wide enough for every sport to make its case."

On the basis of the corpus data, the plural forms of *criterion* are heavily dominated by the use of the foreign plural. The less frequent plural forms ending in *-s* are more marginal than the singular use of *criteria*. Keeping in mind Peters' (2004: 133) observation in Section 5.3.4, the foreign plural form is without question the dominant form of the word. When compared to the other plural forms, even the singular use of *criteria* is significantly more frequent.

7.4 Plural forms of *phenomenon*

The plural forms of *phenomenon* resemble those of *criterion* in that there are two ending in *-s* which cannot be classified in terms of plural versus singular. There is, however, a semantic division too suggested by the literary sources, especially as concerns *phenomenons*. The line between 'something observable' and 'something exceptional' is a matter of the analyst's subjective judgment

[8] Liu, Hua, Qihao Weng and David Gaines. 2008. "Spatio-temporal analysis of the relationship between WNV dissemination and environmental variables in Indianapolis, USA". *International Journal of Health Geographics* 7:66.

in some cases. As illustrated in Table 3 (Section 6.3.1), GloWbE corpus contains tokens in six different plural forms or spellings of the word.

7.4.1 *Phenomena* in BrE

To clarify the classification, categories A3 and B3 are for tokens whose grammatical number is unclear. Category D denotes an unclear semantic division between A and B, and E is for tokens in uninterpretable contexts. The classification as a whole including empty categories is shown below.

Table 15. Classification and token distribution of *phenomena* in BrE

Classification		**Number of tokens out of 150**	**Percentage**
A. Something observable	1. Plural	131	87.33%
	2. Singular	18	12%
	3. Unclear		
B. Something exceptional	1. Plural		
	2. Singular		
	3. Unclear		
C. Proper noun		1	0.67%
D. Multiple/overlapping			
E. Unclear			

It is notable that no tokens could be placed in category B, at least not in the sense ‘a marvel’ or ‘outstanding person’ (see Section 5.4.5). Perhaps the situation could be demonstrated with the phrase: “miraculous *phenomena* like northern lights”. Arguably these natural light displays in the polar regions of Earth are ‘exceptional’. However, in this classification observable natural *phenomena* belong to category A1 and indeed the phrase “natural *phenomena*” is often found in the corpus data.

The use of the foreign plural as a singular form is proportionally more frequent than is the case with *criteria* (Sections 7.3.1 and 7.3.2). For example, token 142:

> Where your argument turns into complete bollocks is the moment you assume without any proof what so ever that climate change is a man made *phenomena*.

As in the above, quite a few of the tokens in the singular can be found in online newspapers' comment sections, which indicates colloquial rather than formal language use. The one proper noun token apparently refers to a piece of hi-fi audio equipment.

7.4.2 *Phenomena* in AmE

Similarly to BrE, no tokens could be allocated to category B. There is, however, also a difference.

Table 16. Classification and token distribution of *phenomena* in AmE

Classification		**Number of tokens out of 150**	**Percentage**
A. Something observable	1. Plural	117	78%
	2. Singular	30	20%
	3. Unclear	3	2%
B. Something exceptional	1. Plural		
	2. Singular		
	3. Unclear		
C. Proper noun			
D. Multiple/overlapping			
E. Unclear			

Table 16 suggests that the use of the foreign plural as a singular form is noticeably more frequent in AmE than BrE, representing one fifth of the analyzed tokens. There is also indication that the usage is not restricted to informal contexts, such as web site comment sections, as the following text example (token 126) by a New York University philosophy professor illustrates:

> But there is one misleading implication involved in calling the *phenomena* I describe as a " shifted spectrum ", namely that there is no reason to believe that there is any sort of uniform displacement of the color wheel, a mini version of the traditional inverted spectrum.

Based on the corpus data, *phenomena* is more frequently used as a singular form than *criteria* and in both cases the difference between the two language varieties is that the singular usage is more common in AmE. Three tokens were inconsistent or ambiguous in their grammatical number.

7.4.3 *Phenomenons* in BrE

While the search word *phenomena* retrieved thousands of tokens in both varieties (Table 3 in Section 6.3.1), there are fewer than forty tokens of *phenomenons* for each. After discarding the tokens that had the same web source, the number is reduced from 39 to 34 in BrE. With the regular *-s* ending, there is no need to classify the grammatical number in the table below.

Table 17. Classification and token distribution of *phenomenons* in BrE

Classification	Number of tokens out of 34	Percentage
A. Something observable	21	61.8%
B. Something exceptional	6	17.6%
D. Multiple/overlapping	7	20.6%

The relatively low number of tokens may not allow far-reaching conclusions but at least two observations call for attention here. Firstly, the fairly broad sense represented by category A is the most frequent use of the regular plural form. Secondly, a peculiar similarity can be found between a group of tokens, which leads to them being placed in category D.

There are seven tokens which refer to recent developments on the internet, computer technology or social networking. For example, token 4:

> The *phenomenons* of Facebook, Twitter, You Tube etc suggest that community is being redefined in a virtual way, but may not provide for appropriate social interaction.

The stereotypical category B representative would be something closer to token 6:

> Couto granted an interview to the Portuguese newspaper " Record " and identified Cristiano Ronaldo as one of the biggest *phenomenons* in football history.

It almost seems as if there is a degree of "exceptionality" that is applied to these new technological *phenomena* in a similar fashion to successful artists or sports stars being sometimes referred to as *phenomenons*. This is of course a view that can be contested and another analyst might place the seven tokens of category D in A as simply describing observable developments.

Overwhelmingly, when the lexeme *phenomenon* is used to describe things occurring in nature, the foreign plural form is preferred. Nevertheless, some exceptions can be found even in online newspapers, as a caption text in a Daily Mail article (token 17) demonstrates:

> Colourful: Several different striking hues are on display in one of nature's greatest *phenomenons* - including grey, blue and red (left), and orange and purple (right)

In this classification, the example above, which describes the event of lava from a volcano flowing into the ocean, belongs to category A.

7.4.4 *Phenomenons* in AmE

After rejecting multiple tokens from the same source, the total number in this case is 31 instead of the 34 in the original search. The regular plural is quite infrequent in both varieties.

Table 18. Classification and token distribution of *phenomenons* in AmE

Classification	Number of tokens out of 31	Percentage
A. Something observable	24	77.4%
B. Something exceptional	5	16.1%
D. Multiple/overlapping	2	6.5%

There are fewer examples of category D, described in the previous Section 7.4.3, and the majority of the tokens fall quite effortlessly into category A. Considering the total number of tokens in B, the sense is remarkably infrequent as opposed to what might be expected on random web sites that also discuss pop culture, sports and so on. It is puzzling why more tokens like the following (token 3) do not occur in the corpus data:

> It's not that there aren't ten movies that are worthy of being on the list, it's just a matter of finding where the gap lies between the great films of the year and the *phenomenons* of the year.

The overall impression is that all the tokens in all the categories are found in more or less informal web sites. There are, for instance, no major news websites or educational institutions among the web sources.

A similar observation to that made in Section 7.3.3 about the forms of *criterion* can be made about *phenomenon*. If the singular use of *phenomena* is proportionally between 12 and 20 percent of the tokens, as the first 150 analyzed instances suggested in Tables 15 and 16, it means that among all the tokens there should be more than 600 such tokens. This in turn implies that the regular use of

phenomenons is less than 1/20 of the frequency of the singular use of *phenomena* alone. In other words, not quite as rare as *criterions* versus *criteria* but very rare nonetheless.

7.4.5 Less frequent plural forms of *phenomenon*

There are even less frequent, very marginal forms found in the corpus data. Table 19 below presents the distortion-corrected number of tokens in both varieties, which adds up to only 20.

Table 19. Distribution of less frequent plural forms of *phenomenon* in BrE and AmE

Search word	Number of tokens in BrE	Number of tokens in AmE
Phenomenas	3	1
Phaenomena	8	5
Phainomena	1	1
Phoenomena	1	0

Given the fact that the analogous plural *criterias* had a combined total of 22, and that the OED listed *phenomenas* as a less frequent plural form, it is interesting that the corpus search only returned a total of four tokens of *phenomenas*. Two of these four (token 2 BrE and token 1 AmE) occur within text passages that include spelling errors. The third is found in a private blog and the fourth in a more mainstream technology website (BrE, token 3):

> Despite it being one of the most common *phenomenas* in the exercise world (particularly common in runners and swimmers) there is little consensus as to what precisely it is and what causes it.

Of these less frequent plural forms, the most frequent in the corpus data is not *phenomenas* but *phaenomena*. A closer inspection reveals the reason for this. Seven out of the thirteen tokens (8 BrE + 5 AmE) refer to the title of an ancient literary work, the poem *Phaenomena* by the Greek poet Aratus and therefore belong to the proper noun category C. The other *phaenomena* tokens seem to be found in quotes or reproductions of older texts, such as Adam Smith's *Theory of Moral Sentiments* from 1759 (token 3, AmE):

> An author who treats of natural philosophy, and pretends to assign the causes of the great *phaenomena* of the universe, pretends to give an account of the affairs of a very distant country, concerning which he may tell us what he pleases, and as long as his narration keeps within the bounds of seeming possibility, he need not despair of gaining our belief.

The spelling *phainomena* with its two tokens is similar in that it occurs once in the title of an ancient Greek work (token 1, BrE) and a second time as a transliteration of terminology used by Socrates (token 1, AmE). The semantic classification of the less frequent plural forms in both varieties combined is presented in the table below, without percentages due to the small total number:

Table 20. Classification of less frequent plural forms of *phenomenon* in BrE and AmE

Classification		**Number of tokens out of 20**
A. Something observable	1. Plural	*Phenomenas* 2 (1 BrE, 1 AmE) *Phaenomena* 6 (2 BrE, 4 AmE) *Phainomena* 1 (AmE)
	2. Singular	*Phoenomena* 1 (BrE)
	3. Unclear	
B. Something exceptional	1. Plural	
	2. Singular	
	3. Unclear	
C. Proper noun		*Phaenomena* 7 (6 BrE, 1 AmE) *Phainomena* 1 (BrE)
D. Multiple/overlapping		*Phenomenas* 1 (BrE)
E. Unclear		*Phenomenas* 1 (BrE)

The only token with the rare spelling *phoenomena* does not occur in a reference to ancient works, nor is it found in an old text but in a newspaper article comment section (token 1, BrE): "If you spent any time in NE Asia this would be a common *phoenomena*." It is also the only instance of these less frequent plural forms being used as a singular form. The less frequent forms of *phenomenon* are extremely rare in the corpus data. Before distortion correction, the combined search results of both varieties amount to 35 compared to the 7994 tokens of *phenomena*. After the correction, the number is lowered to 20.

8. Discussion

This section is dedicated to further discussion on the observations and issues that emerged during the corpus analysis. I will consider possible explanations for some of the observations. The unexpected deficiencies encountered in the use of the GloWbE corpus also deserve more detailed examination, as there are issues I have not yet seen discussed by any other study that has employed the corpus as a tool for analysis.

As pointed out in the introduction in Section 1, my aim was not to test a hypothesis but to compare the corpus data with the literary sources. That is not to say that I did not have any expectations or preconceptions about the corpus data. Some of the findings do indeed go against the intuitive impressions I had formed, drawn from the usage guides and dictionaries. However, in many cases the literary sources and the corpus data are consistent with each other with no real surprises.

The plural forms of the four nouns studied are a diverse group with different factors determining their preferred use. Some of the plural forms returned very low numbers in the GloWbE search results, which means that the most important observation about them is their infrequent occurrence rather than any semantic or other differentiation.

One prominent observation is that the two language varieties exhibit rather small differences, as regards the analyzed tokens' semantic distribution. There are, however, some. The corpus data indicates that the figurative use of *antennae* is approximately twice as common in BrE as in AmE, representing about a third of the instances in BrE. Is that a sign of something idiomatically British manifesting itself in the corpus data? Perhaps the web content selected by the GloWbE sampling frame consists of such BrE web sites where the figurative use is a common part of the register — for example, the intuition or alertness of politicians discussed in newspapers. In any case it seems that the explanation is not a random overrepresentation within the first 150 analyzable tokens, because the distortion-correction of discarding multiple tokens from the same

source causes the last analyzed *antennae* in BrE to reach up to number 206 (see Appendix A) out of the total 236, which means a good representation of the majority of the tokens.

Due to the very large total number of analyzable tokens, I did not classify and quantify all their web sources, which undoubtedly would have enabled a deeper insight into the question of what kind of context is connected to which plural forms.

Another kind of difference between BrE and AmE is to be seen in Table 3 (Section 6.3.1). The total number tokens in BrE and AmE differs significantly as regards *formulae* and *formulas*. The former is more than twice as frequent in BrE as in AmE (599 vs. 265) and the reverse is true with the latter: *formulas* has nearly twice the number of tokens in AmE compared to BrE (1313 vs. 697). This distribution stands out clearly from the other search words of this study, yet it is only commented upon by Peters (2007: 217) in the literary sources, apparently because Peters' guide uses corpus data from two different corpora as the major source of information.

The American preference for the regular form is perhaps a symptom of an ongoing regularization process that either concerns the noun *formula* as a whole or some senses of it. The corpus data indicates that AmE uses *formulas* in the scientific sense more often than BrE (ca. 59% vs. 49% of the tokens), which can be seen as a piece of evidence in support of the regularization assumption. Furthermore, according to Collins (2015: 337), many authors share the view that AmE has a greater tendency towards regularization and colloquialization than BrE when it comes to grammar, which would imply observable differences in inflectional endings too.

The third difference between the two varieties has to do with plural forms being used as a singular, which concerns both *criteria* and *phenomena*. The reinterpretation of plural forms as singular is more frequent in AmE, in the case of *phenomena* almost twice as frequent (20% vs. 12% of the tokens). The literary sources provided no specific information on the frequency of such use. However, Peters (2004: 134) gives an interesting indication of a potentially very high frequency among some groups of users in an informal register:

> *Criteria* not uncommonly serves for the singular in conversation, and in research among young Australian adults by Collins (1979), more than 85% treated it as a singular.

Certainly 20% of the tokens is a proportion that deserves recognition and calls for some explanation. Some literary sources suggested that there might be a development going on that is taking these *-a* ending plural forms towards an acceptable status as singular forms (e.g. Section 5.4.4) in the wake of *data*, *media* or *agenda*. Peters (ibid.) uses the term "collective or singular noun". A development towards a singular form may be aided by the very lack of the usual *-(e)s* ending itself, as well as the decline of the knowledge of classical languages mentioned by Burchfield (1996: 442).

The collectiveness of *data* and *media* seems quite natural in the sense that they often refer to a group that consists of similar or comparable sub-entities. It would seem very odd to itemize the constituents of, for instance, digital data, especially since it all practically consists of millions or billions of zeros and ones.

Criteria is similar in the way that it is almost always used to refer to a group of requirements that 'come in a bundle', as it were. There hardly ever is just one *criterion*. Perhaps these real world conditions that make the (originally) plural form so overwhelmingly more common than the standard singular form contribute to the singular form eventually vanishing from use.

While it is plausible that *criteria* is on the way of becoming an accepted collective noun just like *data*, due to the lack of the need to differentiate between plural and singular, I would argue that such factors affect *phenomena* less. Despite the form occurring frequently in phrases like "natural *phenomena*" or "observable *phenomena*", the very broad sense of the word allows it to describe almost any one event or occurrence. I would claim that this maintains the need to preserve separate singular and plural forms, at least if any kind of iconicity (see Section 4.5) or clarity is sought. There will continue to be a need to be able to say, for example "this is a strange *phenomenon*". Regardless of this speculation, the fact on the ground (i.e. in the corpus data) is that *phenomena* is

used as a singular form with every fifth token in AmE. So, what else could be at play here? The following is a suggestion.

Let us first consider the following citation taken from the OED entry for *phenomenon*:

> Pronunciation:
>
> Brit. /fɪˈnɒmɪnən/, U.S. /fəˈnɑməˌnɑn/, /fəˈnɑmənən/

More specifically, let us consider the first given AmE pronunciation /fəˈnɑməˌnɑn/. As pointed out by Dimitrova (2010: 2-8), the General American pronunciation often has back unrounded vowels where the Standard British would prefer rounded vowels. In the example word *phenomenon*, the back unrounded /ɑ/ occurs twice. According to McMahon (1994: 72), final nasal consonants seem to be cross-linguistically unstable during the process of language change. If we remove the final nasal from the AmE *phenomenon* the result is /fəˈnɑməˌnɑ/, essentially the foreign plural form of the word. Thus, I would argue that instead of the process of becoming a collective singular noun, the singular use of *criteria* (to lesser extent) and especially of *phenomena* is driven by phonological motivations, particularly AmE vowel qualities combined with the loss of the final /n/.

As regards the regular plural forms *criterions* and *phenomenons*, they are completely overshadowed by their foreign plural counterparts, even the singular use of *criteria* and *phenomena* alone by about 100:1 and 20:1 (Sections 7.3.3 and 7.4.4). This would indicate that there is no significant process of regularization by analogical extension going on with the plural forms of the two nouns. If what I suggested above is true, then the two nouns would demonstrate a case where phonological convenience overrides the advantages of morphological clarity and consistency and hold back the possible tendency to regularize the plural form.

The very infrequent occurrence of especially *phenomenons* in the corpus data is probably the most unexpected observation for me. After all, it was recognized by many of the literary sources as a legitimate plural form with its own separate sense from *phenomena*, a sense that might as well have occurred more among the internet sources. Despite the low number of the tokens of the regular

plural, I will return to the observation made in Section 7.4.3. and elaborate on the issue of the form referring to *phenomenons* of the internet age.

I would propose two possible reasons for the occurrence of phrases such as "the *phenomenons* of Facebook, Twitter, Youtube etc.." or "social media is one of the great *phenomenons* of our age". Firstly, as already mentioned in Section 7.4.3, there could be a semantic component, a connotation which makes the newly emerged and rapidly expanded internet *phenomenons* in a way 'celebrity-like'. They are likened to other cultural exceptional occurrences that in my analysis formed category B 'something exceptional'. For example, suddenly successful popstars would sometimes be called *phenomenons*. Secondly, the fairly recent emergence of these internet 'wonders' might itself encourage the use of the regular plural. As an analogy, the plural for *mouse* in the sense *computer mouse* is not only *mice* but increasingly *mouses* (Huddleston et al. 2002: 1590). This is not a claim I make with certainty but a speculation on the persistence and even recent adoption of the regular plural in a situation where *phenomena* is unquestionably the dominant form. The fact that the two plural forms have co-existed for hundreds of years and the foreign plural still has such a strong numerical representation would imply that no large scale regularization is going to happen anytime soon.

Returning to the more frequent plural forms and the noun *antenna*, it can be pointed out that the literary sources and the corpus data are in harmony on several points. Those literary sources that only provided a general guideline did it along the main semantic division of *antennae* for insects and figurative use and *antennas* for devices. Others that were more detailed recognized that the foreign plural is also used to refer to devices.

It seems that the figurative use is very closely tied to the foreign plural. Of course, it is understandable that it has emerged as a metaphor of the *antennae* in the natural world, but the fact that the figurative use is as rare as between 1% and 4% for *antennas* is very interesting.

In the corpus data, *antennae* is by no means restricted to zoological referents or figurative usage as around a third of the tokens referred to technical devices. This is quite close to the detailed information given by Garner cited in Section 5.3.2, who suggested a quarter, and also indicative of the fact that the less detailed guideline of "*antennae* for insects and *antennas* for devices" does not reflect the actual usage accurately enough.

The regular plural *antennas* is semantically very uniform with 90% - 96% of the tokens having a device as referent, a figure also predicted by Garner. In a way *antennas* is an example of a 'unique and constant' grammatical marker discussed in Section 4.5, although the uniqueness is eroded by the fact that *antennae* can have the same referent and quite often does. The combined semantic distribution of both plural forms of *antenna* in both varieties is shown below.

Table 21. Combined semantic distribution of *antennae* and *antennas* in BrE and AmE

Classification	***Antennae* distribution percentages**	***Antennas* distribution percentages**
A. Zoology	37%	3.33%
B. Device	36.33%	93%
C. Figurative	22.33%	2.67%
D. Proper noun	1.67%	0.67%
E. Multiple/overlapping	1.67%	0.33%
F. Unclear	1%	0%
	100%	100%

Apart from the more frequent figurative use of *antennae* in BrE and the more frequent scientific use of *formulas* in AmE, the corpus data portrays the analyzed tokens of the plural forms of *antenna* and *formula* quite similarly in the two varieties.

Some literary sources provided semantic distribution data on the different senses or uses of *antennae* and *antennas* and that data turned out to be consistent with the analyzed corpus data. With the plural forms of *formula*, the literary sources did not cite any distribution figures. The usual information given was the general advice along the lines of "*formulae* especially for scientific use,

formulas for general use". Furthermore, there was great variation in the detail of semantic definitions, i.e. the number of senses attributed to the noun. In this sense, the corpus data analysis could provide more thorough information than was otherwise available.

As already discussed earlier in this section, there is evidence of AmE preferring the regular plural in the case of *formula* compared to BrE, which may be due to an ongoing regularization process. This study was not designed to examine diachronic data and therefore can only offer conjectures on the issue. The detailed information that can be given is the semantic distribution. When both varieties are combined the distribution is the following:

Table 22. Combined semantic distribution of *formulae* and *formulas* in BrE and AmE

Classification	***Formulae* distribution percentages**	***Formulas* distribution percentages**
A. Scientific	72%	53.67%
B. Method	12%	28.33%
C. Fixed set of words	10%	7.67%
D. Ingestible substance	1.67%	7.67%
E. Motor racing	3.67%	2%
F. Multiple/overlapping	0.33%	0%
G. Unclear	0.33%	0.33%
H. Proper noun	0%	0.33%
	100%	100%

The advice given in the literary sources is consistent with the corpus data in that *formulae* occurs especially, but not only, in scientific use. The sources did not indicate that the regular plural would occur to such an extent with scientific referents, nor did they provide predictions on the frequency of the other senses of the noun. In fact, one of the dictionaries grouped categories A and B above as one. The corpus data suggests that the regular plural is more common when referring to a method or infant formula, but otherwise the figures are not that clear.

In the remaining part of this section, I will discuss the observations that relate to the GloWbE corpus itself. This is because there are deficiencies that unfamiliar users ought to be aware of when setting out to do an analysis using the corpus data.

The first deficiency was introduced in Section 6.3.2. The GloWbE user interface does not treat web sources in such a way that that it would exclude multiple tokens from the same source. With the time and resources at hand, I could not explore why such an obvious flaw is allowed to remain. Common sense would conclude that a digital user interface that automatically discards multiple tokens from the same text would be extremely easy to design. As long as such a system does not exist the process of manual analysis remains unnecessarily complicated. I will illustrate this with the following example. The GloWbE 'context' view displays two similar web addresses (consecutively or close by), e.g. in the form *guardian.co.uk*. The analyst has no way of knowing right away whether the link to the source leads to the very same text where the search word occurs multiple times, or whether the search words occur on different web pages, e.g. different articles under the same web page/domain of the Guardian online newspaper.

In addition, the exact same text passage containing the same token can occur more than once in the search results, not only next to the previous one but later on the list. For example token 40 (*formulae*, BrE) occurs later as token 146. The web links are obsolete but from the address information it can be seen that this is the same text occurring on two different web sites under the same web page *debate.org.uk*. Such instances were naturally excluded from the analysis when noticed, but with almost 1900 tokens to analyze, it was impossible to keep in mind every encountered text - something that digital processing would do very easily.

A second deficiency, referred to at the end of Section 6.3.4, interferes with falsifiability and replicability. Due to the subjective human component of the semantic, qualitative analysis in my thesis, the analysis must be open to debate. The only way to expose debatable points is to make the analysis replicable and falsifiable. This means that any other analyst must be able to trace the tokens

I have analyzed and connect them to the classification I have used. This connection is provided at the end of the study in Appendices A – D. However, during the analysis it turned out that a token may occur with a different search result number in a later search than it did previously. For instance, token 188 (*formulae*, BrE) occurred as token 188 on one search but 187 on another. This curiosity was only noticed because the text passage was saved for possible citation purposes and re-checked. I did not anticipate such a flaw and do not have a systematic record of how common it is. What I can say is that all such instances that I came across had a deviation of one number, not more. Nevertheless, such an issue should not exist complicating falsifiability and replicability.

A third GloWbE deficiency has to do with how the source text is reproduced in the 'expanded context' view. This view displays a few sentences of the source text where the search word is found. As there are many dysfunctional web links to the original web sources (many web sites have ceased to exist), the context view is often the only possible way to examine the context of the token and make conclusions about it. By accident, I encountered an instance where the expanded context text does not match with that of the original source, which still had a functional link. Token 75 (*antennas*, BrE) is taken from an old book where the plural form *antennae* occurs several times on one page[9]. The page is in the form of a picture on that web site and apparently this causes the misinterpretation of *-e* into *-s* by the GloWbE corpus. Thus, the corpus search displays the multiple *antennae* tokens of the original source as *antennas* on the search result list and expanded context view. I would suspect this is not a frequent problem, but with many dysfunctional original web links, it is impossible to be certain.

[9] http://biostor.org/reference/60073/page/3

9. Conclusion

In my thesis, I have examined the plural forms of *antenna*, *formula*, *criterion* and *phenomenon* in two varieties of English, British and American. I have analyzed 1885 tokens using electronic corpus data from the GloWbE corpus with guidance from literary sources such as grammars, language usage guides and dictionaries. I had chosen a corpus-based, descriptive approach and I set out to do the task by posing three research questions.

The first research question was to find out the distribution of the plural forms of *antenna*, *formula*, *criterion* and *phenomenon* in terms of semantics and between British and American English. This question has elaborate answers that have essentially been given in the two previous sections, for the most part in the corpus data analysis Section 7. The corpus data shows that these four nouns all have their particular characteristics. In terms of frequency and semantics, the most common occurrence of the plural forms of *antenna* is the regular plural *antennas* referring to technical devices. It is also the most semantically uniform of all the plural forms that were semantically classified. In the case of *formula*, the most frequent is the foreign plural *formulae* referring to sequences of symbols used in calculation.

The plural forms of *criterion* and *phenomenon* are more numerous but the less frequent forms occur very rarely and these nouns are semantically not as diverse as *antenna* and *formula*. Both *criterion* and *phenomenon* exhibit a noticeable use of their foreign plural forms as singular forms in English, which in my own reasoning is affected by phonological conditions and according to literature, at least for *criteria*, a development into a collective singular noun. The differences between British and American English are often very small. The most noticeable differences are the British tendency to use *antennae* figuratively more, and the American tendency to use regular plural *formulas* more, as well as the more frequent American singular use of *criteria* and *phenomena*.

The second research question asked to what extent the literary sources agree with the corpus data. The answer can be condensed into the following: the more detailed the literary source is in its

description, the better it reflects the corpus data. The sources that are informed by data from other corpora also predict the distributions in GloWbE very accurately. Most of the literary sources do not present numerical distribution data. Although the basic information proved accurate, the descriptions were often short and therefore left open questions. For example, the frequency of *antennae* being used to refer to technical devices was greater than implied by some of the literary sources. The sources also did not explicitly indicate that the use of *phenomena* as a singular could be as frequent as one fifth of the analyzed tokens in American English.

The third research question asked for any other relevant observations arising from the analysis. Perhaps the very rare occurrence of the less frequent plural forms of *phenomenon* and *criterion* is such an observation. The most prominent one is, however, the handful of deficiencies that the GloWbE corpus revealed during the analysis. They relate to the inability of the corpus to exclude multiple samples form the same source, the occasionally incoherent search result numbering and an apparent problem in the interpretation of some source texts correctly onto the expanded context view.

There are different aspects that were excluded from this study which could provide important information on the plural forms in question. A detailed analysis of the contexts the plural forms occur in, for example, or a diachronic perspective on possible changes taking place, or a research into other varieties than the two dealt with in this study. Furthermore, a study of the collocates of these plural forms would certainly yield further relevant information. These are just some examples of the possibilities offered by corpus linguistic methods. I would predict that the use of corpus-informed grammars, usage guides and dictionaries will only increase in the future.

References

Corpus of Global Web-Based English (GloWbE)

https://www.english-corpora.org/glowbe/

Primary literary sources

Grammars

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Declerck, Renaat. 1992. *A Comprehensive Descriptive Grammar of English*. First published 1991. Tokyo: Kaitakusha.

Huddleston, Rodney, Geoffrey K. Pullum, and Laurie Bauer. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Leech, Geoffrey N. and Jan Svartvik. 2002. *A Communicative Grammar of English*. 3rd ed. London: Longman.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Usage Guides

Burchfield, R.W. ed. 1996. *The New Fowler's* Modern English Usage. 3rd edition of Fowler's *A Dictionary of Modern English Usage*. Oxford: Oxford University Press.

Crystal David, ed. 2009. *H.W. Fowler* A Dictionary of Modern English Usage *the Classic First Edition with a new Introduction & Notes by David Crystal*. Oxford: Oxford University Press.

Davidson, Mark. 2001. *Watchwords: A Dictionary of What's Right, Wrong, & Risky in Today's American English Usage*. South Pasadena: Harrison.

Garner, Bryan. 2003. *Garner's Modern American Usage*. 2nd edition. Oxford: Oxford University Press.

Garner, Bryan. 2016. *Garner's Modern English Usage*. 4th edition. Oxford: Oxford University Press.

Howard, Godfrey. 1993. *The Guide to Good English: English Usage in the 1990s*. London: Macmillan.

Peters, Pam. 2004. *The Cambridge Guide to English Usage*. Cambridge: Cambridge University Press.

Swan, Michael. 2005. *Practical English Usage*. 3rd edition. Oxford: Oxford University Press.

Dictionaries

Deuter, Margaret, Jennifer Bradberry, and Joanna Turnbull, eds. 2015. *Oxford Advanced Learner's Dictionary*. 9th edition. Oxford: Oxford University Press.

Gove, Philip B, ed. 1961. *Webster's Third New International Dictionary*. Springfield, MA: Merriam-Webster Inc.

Hellweg, Paul, ed. 2007. *The American Heritage College Dictionary*. 4th edition. Boston: Houghton Mifflin.

McIntosh, Colin, ed. 2013. *Cambridge Advanced Learners Dictionary*. 4th edition. Cambridge: Cambridge University Press.

Mish, Frederick C, ed. 2003. *Merriam-Webster's Collegiate Dictionary*. 11th edition. Springfield, MA: Merriam-Webster Inc.

OED Online. Oxford: Oxford University Press. Accessed at https://www-oed-com.libproxy.tuni.fi/

Pickett, Joseph P, ed. 2011. *The American Heritage Dictionary of the English Language*. 5th edition. Boston: Houghton Mifflin.

Sinclair, John, et al., eds. 2014. *Collins COBUILD Advanced Learners Dictionary*. 8th edition. Glasgow: HarperCollins.

Sparks, Andrew N, Jonathan L. Goldman, and Donald Steward, eds. 2014. *Webster's New World College Dictionary*. 5th edition. Boston: Houghton Mifflin Harcourt.

Stevenson, Angus, ed. 2010. *The Oxford Dictionary of English*. 3rd ed. Oxford: Oxford University Press.

Other Works Cited

Baker, Peter S. 2012. *Introduction to Old English*. 3rd edition. Chichester, UK: Wiley-Blackwell.

Ball, Carleton R. 1928. "English or Latin Plurals for Anglicized Latin Nouns?" *American Speech* 3: 291-325.

Collins, Peter. 2015. *Grammatical Change in English World-wide*. Amsterdam: John Benjamins.

Dimitrova, Snezhina. 2010. *British and American Pronunciation*. New York: Oxford University Press.

Durkin, Philip. 2014. *Borrowed Words: A History of Loanwords in English*. Oxford: Oxford University Press.

Ferdinand, Vanessa, Simon Kirby, and Kenny Smith. 2019. "The Cognitive Roots of Regularization in Language." *Cognition* 184: 53-68.

Fischer, Olga, Ans van Kemenade, Willem Koopman, and Wim van der Wurff. 2001. *The Syntax of Early English*. Cambridge: Cambridge University Press.

van Gelderen, Elly. 2014. *History of the English Language*. Revised edition. John Benjamins.

Görlach, Manfred. 1991. *Introduction to Early Modern English*. Cambridge: Cambridge University Press.

Jacobs, Rachel. 2009. *Latin Grammar: A QuickStudy Language Reference Guide*. Boca Raton, Florida: BarCharts Inc.

Jespersen, Otto. 1912. *Growth and Structure of the English Language*. 2nd ed. Leipzig: B.G. Teubner.

Marshall, Chloë R. and Heather K. J. van der Lely. 2012. “Irregular past tense forms in English: how data from children with specific language impairment contribute to models of morphology.” *Morphology* 22: 121-141.

McCarthy, Michael. Foreword to *Corpus-based Research in Applied Linguistics: Studies in Honor of Doug Biber*, edited by Cortes, Viviana and Eniko Csomay. 2015. Amsterdam: John Benjamins.

McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Methods, Theory and Practice*. Cambridge: Cambridge University Press.

McMahon, April M. S. 1994. *Understanding Language Change*. Cambridge: Cambridge University Press.

Smyth, Herbert W. 1956. *Greek Grammar*. Revised by Gordon M. Messing. Cambridge, Mass.: Harvard University Press.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Vennemann, Theo. 1978. *Phonetic and Conceptual Analogy*. In *Recent Developments in Historical Phonology*, edited by Baldi, P. and Werth, R. N. 341-408. The Hague: Mouton.

Warfelt, Leslie M, ed. 2012. *Language Acquisition*. Hauppauge, N.Y.: Nova Science Publishers, Inc.

Appendix A. Classification and token numbers of the plural forms of *antenna*

Search word, language variety and classification	**Token numbers in GloWbE search results**
Antennae, BrE, A. Zoology	2, 8, 9, 10, 12, 13, 15, 18, 22, 25, 27, 31, 32, 41, 43, 46, 48, 56, 58, 63, 68, 69, 71, 81, 85, 88, 95, 99, 105, 114, 118, 122, 123, 124, 126, 129, 132, 136, 144, 145, 147, 150, 151, 157, 180, 183, 189, 196, 198, 204, 205
Antennae, BrE, B. Device	1, 5, 6, 11, 14, 21, 23, 26, 28, 29, 40, 42, 45, 51, 53, 57, 60, 64, 65, 70, 75, 77, 82, 84, 87, 89, 93, 96, 106, 107, 109, 116, 117, 125, 133, 134, 137, 139, 142, 155, 156, 158, 179, 182, 188, 191, 195, 197, 203, 206
Antennae, BrE, C. Figurative use	4, 16, 17, 20, 24, 30, 44, 47, 49, 50, 52, 54, 55, 62, 67, 72, 74, 76, 78, 80, 86, 92, 94, 104, 108, 115, 119, 120, 121, 127, 128, 130, 131, 138, 141, 148, 149, 152, 154, 181, 186, 187, 192, 199, 201
Antennae, BrE, D. Proper noun	39, 83, 110
Antennae, BrE, E. Multiple/overlapping	184
Antennae, AmE, A. Zoology	2, 3, 4, 5, 6, 7, 8, 9, 12, 13, 17, 19, 20, 22 , 31, 34, 41, 47, 48, 49, 52, 53, 54, 56, 62, 66, 70, 71, 74, 79, 84, 96, 104, 105, 106, 117, 118, 120, 121, 123, 125, 126, 127, 128, 130, 138, 178, 182, 188, 192, 195, 198, 204, 208, 210, 217, 220, 222, 225, 226
Antennae, AmE, B. Device (**s**) = used as a singular (**m**) = misspelling	1, 10, 16, 18, 23, 25, 29, 32, 38, 40, 57, 58, 59, 60, 61, 63, 64(**s**), 65, 67(**s**), 75, 78(**s**), 80, 83, 95, 98, 99, 101, 103, 107(**m**), 109, 115, 116, 119, 122, 124, 135, 137, 139, 143, 177, 186, 190, 191, 196, 199, 200, 201, 202, 203, 207, 211, 214, 216, 218, 219, 227, 229, 233, 234
Antennae, AmE, C. Figurative use (**s**) = used as a singular	24, 30, 33, 50, 51, 68, 69, 72, 86, 97, 100, 114, 136, 142, 185, 187, 205, 221, 228, 230, 231(**s**), 232
Antennae, AmE, D. Proper noun	112,113
Antennae, AmE, E. Multiple/overlapping (**s**) = used as a singular	21, 82(**s**), 141, 189
Antennae, AmE, F. Unclear	94, 102, 209
Antennas, BrE, A. Zoology	122, 228, 244,
Antennas, BrE, B. Device	1, 2, 5, 13, 15, 17 ,21, 22, 23, 24, 25, 27, 28, 29, 33, 34, 35, 36, 38, 39, 40, 41, 42, 43, 44, 25, 46, 47, 49, 50, 51, 53, 55, 59, 60, 61, 62, 69, 70, 71, 72, 73, 74, 79, 83, 84, 85, 88, 89, 90,91, 92, 93, 94, 95, 96, 102, 103, 104, 108, 111, 115, 116, 120, 121, 123, 124, 126, 134, 136, 138, 139, 140, 142, 143, 144, 147, 148, 149, 150, 151, 152, 154, 155, 158, 159, 161, 162, 163, 164, 165, 167, 168, 169, 171, 172, 174, 175, 176, 177, 178, 179, 181, 182, 183, 184, 185, 196, 198, 199, 200, 201, 203, 204, 205, 208, 209, 212, 214, 215, 216, 222, 223, 225, 227, 229, 230, 231, 233, 234, 236, 237, 238, 239, 240, 243, 245, 246, 248, 251, 252, 254, 257, 258
Antennas, BrE, C. Figurative use	166, 259
Antennas, BrE, D. Proper noun	146
Antennas, AmE, A. Zoology	3, 49, 84, 86, 106, 168, 245
Antennas, AmE, B. Device	1, 2, 6, 7, 8, 9, 10, 11, 12, 17, 18, 21, 22, 23, 24, 25, 27, 35, 37, 42, 43, 44, 45, 46, 48, 50, 51, 53, 54, 56, 57, 58, 62, 63, 65, 66, 68, 69, 71, 73, 74, 76, 79, 81, 83, 87, 88, 96, 98, 100, 107, 109, 110, 111, 112, 113, 114, 117, 118, 121, 122, 123, 128, 129, 130, 132, 133, 135, 136, 137, 138, 139, 140, 143, 152, 153, 154, 156, 157, 158, 159, 160, 164, 165, 170, 171, 172, 173, 174, 176, 177, 178, 180, 181, 184, 185, 186, 187, 188, 192, 193, 195, 196, 197, 198, 201, 204, 205, 206, 208, 209, 211, 212, 213, 217, 218, 220, 221, 222, 223, 224, 227, 229, 233, 236, 237, 238, 241, 243, 244, 246, 250, 253, 254, 257
Antennas, AmE, C. Figurative use	61, 67, 183, 207, 242, 248
Antennas, AmE, D. Proper noun	191
Antennas, AmE, E. Multiple/overlapping	36

Appendix B. Classification and token numbers of the plural forms of *formula*

Search word, language variety and classification	**Token numbers in GloWbE search results**
Formulae, BrE, A. Scientific	2, 4, 5, 6, 9, 13, 14, 16, 18, 21, 23, 24, 25, 26, 27, 29, 33, 34, 35, 47, 48, 49, 59, 60, 64, 67, 69, 71, 72, 73, 74, 75, 84, 85, 87, 88, 91, 96, 97, 98, 100, 101, 109, 112, 114, 115, 119, 120, 121, 123, 124, 126, 127, 128, 129, 130, 131, 132, 133, 136, 141, 142, 157, 158, 159, 160, 161, 162, 168, 172, 175, 177, 178, 182, 183, 185, 187, 189, 191, 192, 193, 196, 197, 198, 204, 207, 208, 212, 214, 215, 217, 218, 219, 220, 222, 223, 224, 225, 226, 227, 228, 230, 232, 233, 236, 239, 241, 242, 246, 247, 248, 250, 252
Formulae, BrE, B. Method	15, 31, 63, 66, 90, 99, 122, 139, 153, 176, 202, 221, 234, 249
Formulae, BrE, C. Fixed set of words	1, 40, 61, 65, 125, 190, 211, 213, 229, 254
Formulae, BrE, D. Ingestible substance	118, 188
Formulae, BrE, E. Motor racing	3, 17, 19, 32, 137, 181, 200, 206, 244, 251
Formulae, BrE, G. Unclear	199
Formulae, AmE, A. Scientific (**s**) = used as a singular	4, 5, 11, 14, 16, 17, 19, 20, 22, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 40, 42, 43, 44, 47, 49, 51, 52, 53, 72, 73, 75, 78, 79, 82, 83, 85, 86, 87, 89, 91, 92, 93, 94, 96, 97, 102, 103, 104, 105, 108, 114, 115, 121, 125, 126, 128, 138, 139, 142(**s**), 143, 144, 145, 147, 148, 149, 150, 152, 153, 158, 159, 160, 162, 164, 165, 166, 167, 169, 175, 179, 184, 186, 187, 191, 193, 194, 195, 199, 200, 205(**s**), 206, 207, 208, 209, 210, 211, 213, 215, 216, 217, 219, 220, 221
Formulae, AmE, B. Method	1, 6, 36, 37, 61, 74, 77, 84, 101, 123, 124, 130, 132, 133, 163, 168, 183, 196, 197, 204, 214, 222
Formulae, AmE, C. Fixed set of words	2, 12, 15, 18, 23, 39, 41, 50, 88, 107, 109, 118, 127, 131, 157, 180, 188, 190, 218, 223
Formulae, AmE, D. Ingestible substance	24, 122, 131
Formulae, AmE, E. Motor racing	202
Formulae, AmE, F. Multiple/overlapping	95
Formulas, BrE, A. Scientific	1, 3, 8, 9, 10, 14, 17, 19, 20, 25, 28, 30, 31, 34, 36, 42, 43, 44, 47, 49, 50, 54, 55, 56, 57, 62, 65, 68, 70, 72, 73, 76, 82, 84, 87, 89, 93, 95, 98, 100, 101, 112, 113, 115, 117, 119, 123, 127, 136, 140, 142, 143, 146, 148, 151, 153, 156, 161, 162, 169, 171, 172, 177, 180, 185, 188, 189, 196, 198, 199, 201, 204, 207
Formulas, BrE, B. Method	2, 11, 12, 13, 15, 18, 26, 27, 32, 35, 39, 41, 52, 63, 64, 74, 75, 80, 81, 83, 88, 105, 116, 118, 121, 125, 126, 128, 130, 132, 133, 134, 135, 147, 160, 170, 174, 186, 187, 191, 192, 194, 195, 206
Formulas, BrE, C. Fixed set of words	40, 86, 97, 106, 114, 124, 129, 130, 137, 139, 150, 152, 197
Formulas, BrE, D. Ingestible substance	38, 48, 66, 67, 96, 122, 131, 138, 175, 178, 193, 200
Formulas, BrE, E. Motor racing	29, 33, 90, 94, 141, 190
Formulas, BrE, G. Unclear	85
Formulas, BrE, H. Proper noun	53
Formulas, AmE, A. Scientific	1, 3, 4, 8, 25, 27, 28, 31, 32, 36, 38, 39, 42, 45, 46, 47, 48, 50, 52, 56, 57, 58, 60, 61, 62, 66, 71, 72, 73, 77, 79, 81, 84, 89, 92, 93, 94, 102, 103, 104, 105, 106, 110, 112, 114, 116, 117, 119, 120, 121, 122, 123, 126, 127, 129, 131, 133, 136, 140, 143, 145, 146, 147, 153, 154, 158, 159, 160, 161, 164, 168, 169, 170, 171, 172, 173, 174, 183, 190, 191, 192, 201, 207, 215, 217, 222, 223, 226
Formulas, AmE, B. Method	2, 5, 7, 14, 16, 17, 18, 24, 41, 63, 75, 83, 86, 87, 90, 107, 128, 130, 132, 141, 144, 148, 149, 162, 176, 177, 178, 180, 182, 188, 194, 195, 203, 206, 214, 218, 219, 220, 221, 224, 225
Formulas, AmE, C. Fixed set of words	9, 15, 21, 30, 64, 65, 74, 97, 108, 142
Formulas, AmE, D. Ingestible substance	6, 10, 22, 40, 59, 70, 82, 88, 179, 189, 193

Appendix C. Classification and token numbers of *criteria*

Search word, language variety and classification	**Token numbers in GloWbE search results**
Criteria, BrE, A. Plural	1, 2, 3, 4, 11, 13, 17, 18, 20, 21, 23, 25, 26, 27, 29, 30, 32, 33, 34, 37, 38, 42, 43, 44, 47, 48, 49, 53, 59, 60, 62, 63, 64, 65, 66, 67, 68, 69, 73, 74, 75, 76, 77, 78, 80, 81, 82, 84, 88, 89, 90, 91, 92, 93, 94, 96, 97, 98, 99, 100, 101, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 118, 119, 120, 124, 130, 131, 132, 135, 136, 137, 138, 139, 140, 143, 145, 146, 147, 148, 149, 151, 152, 153, 154, 155, 156, 158, 159, 160, 162, 163, 164, 166, 170, 172, 172, 175, 177, 191, 192, 193, 194, 196, 197, 208, 209, 210, 216, 217, 218, 219, 220, 221, 231, 232, 236, 243, 251, 252, 253, 254, 255, 256, 257, 258, 260
Criteria, BrE, B. Singular	12, 31, 41, 125, 126, 129, 141, 161, 234
Criteria, BrE, C. Unclear	133
Criteria, BrE, D. Proper noun	22
Criteria, AmE, A. Plural	1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 21, 22, 23, 24, 27, 29, 31, 32, 33, 34, 35, 37, 38, 39, 41, 42, 45, 46, 51, 52, 55, 56, 59, 60, 61, 63, 66, 67, 68, 69, 77, 78, 79, 80, 85, 87, 88, 89, 91, 94, 95, 111, 112, 115, 117, 120, 121, 123, 125, 129, 130, 131, 132, 137, 139, 140, 142, 143, 145, 146, 148, 149, 150, 151, 157, 158, 159, 160, 162, 163, 167, 168, 169, 170, 171, 173, 174, 175, 184, 186, 191, 192, 193, 196, 197, 200, 201, 205, 206, 207, 208, 209, 210, 211, 224, 225, 230, 231, 234, 236, 237, 239, 241, 242, 243, 244, 245, 250, 251, 253, 254, 258, 259, 260, 261, 262, 281, 283, 288, 289
Criteria, AmE, B. Singular	12, 20, 25, 43, 49, 57, 90, 93, 147, 165, 166, 172, 180, 287
Criteria, AmE, C. Unclear	3

Appendix D. Classification and token numbers of the plural forms of *phenomenon*

Search word, language variety and classification	**Token numbers in GloWbE search results**
Phenomena, BrE, A1. Something observable, Plural	1, 2, 3, 4, 6, 7, 10, 11, 12, 15, 17, 18, 23, 26, 27, 28, 29, 30, 35, 36, 38, 39, 40, 41, 42, 45, 46, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57, 59, 60, 61, 62, 65, 66, 67, 68, 69, 78, 79, 82, 84, 88, 90, 91, 92, 98, 99, 100, 103, 105, 106, 107, 108, 111, 112, 113, 115, 117, 118, 119, 123, 124, 126, 127, 128, 129, 130, 134, 135, 136, 137, 140, 141, 143, 144, 147, 148, 150, 155, 156, 157, 160, 168, 169, 170, 171, 172, 175, 176, 177, 179, 180, 182, 183, 185, 186, 187, 188, 192, 194, 196, 198, 199, 200, 202, 203, 206, 208, 209, 210, 211, 212, 214, 216, 2017, 218, 220, 222, 228, 229, 231, 234
Phenomena, BrE, A2. Something observable, Singular	13, 16, 24, 77, 85, 89, 138, 139, 142, 149, 167, 184, 205, 215, 221, 223, 233
Phenomena, BrE, C. Proper noun	37
Phenomena, AmE, A1. Something observable, Plural	2, 4, 5, 7, 9, 11, 12, 13, 14, 15, 16, 17, 27, 28, 42, 51, 52, 54, 55, 56, 57, 58, 60, 62, 65, 70, 72, 73, 75, 76, 77, 81, 93, 96, 98, 100, 101, 103, 104, 105, 111, 112, 116, 118, 119, 121, 123, 125, 127, 128, 129, 135, 136, 137, 150, 151, 152, 154, 158, 160, 162, 163, 164, 165, 168, 169, 171, 175, 176, 178, 179, 180, 194, 195, 196,198, 200, 201, 202. 203, 205, 207, 208, 209,212, 213, 218, 219, 220, 221, 222, 223, 224, 226, 227, 228, 229, 236, 237, 238, 239, 242, 244, 245, 246, 248, 250, 252, 253, 255, 260, 261, 264
Phenomena, AmE, A2. Something observable, Singular	1, 3, 6, 31, 59, 61, 68, 71, 92 ,95, 99, 110, 113, 114, 115, 117, 120, 122, 124, 126, 138, 149, 156, 157, 159, 161, 166, 172, 173, 197, 204, 243, 263
Phenomena, AmE, A3. Something observable, Unclear	25, 29, 199
Phenomenons, BrE, A. Something observable	1, 2, 3, 5, 9, 13, 14, 16, 17, 18, 22, 21, 22, 23, 24, 25, 27, 28, 29, 34, 35, 36, 39
Phenomenons, BrE, B. Something exceptional	6, 12, 19, 30, 31, 33
Phenomenons, BrE, D. Multiple/overlapping	4, 10, 11, 20, 21, 26, 38
Phenomenons, AmE, A. Something observable	1, 2, 4, 7, 8, 9, 10, 12, 13, 14, 15, 20, 21, 22, 23, 24, 26, 27, 29, 30, 31, 32, 33, 34
Phenomenons, AmE, B. Something exceptional	3, 5, 17, 18, 19
Phenomenons, AmE, D. Multiple/overlapping	6, 25
Phenomenas, BrE, A1. Something observable, Plural	3
Phenomenas, BrE, D. Multiple/overlapping	1
Phenomenas, BrE, E. Unclear	2
Phenomenas, AmE, A1. Something observable, Plural	1
Phaenomena, BrE, A1. Something observable, Plural	7, 9
Phaenomena, BrE, C. Proper noun	1, 2, 3, 4, 5, 8
Phaenomena, AmE, A1. Something observable, Plural	1, 2, 16, 17
Phaenomena, AmE, C. Proper noun	18
Phainomena, BrE, C. Proper noun	1
Phainomena, AmE, A1. Something observable, Plural	1
Phoenomena, BrE, A1. Something observable, Singular	1